Gradi di Grammaticalità e Acceptability: La Dissoluzione della Dicotomia Grammaticale nella Teoria Linguistica Contemporanea

Introduzione: La Concezione Tradizionale e i suoi Fondamenti Epistemologici

La concezione tradizionale della grammaticalità, ereditata dal paradigma generativista chomskiano nella sua formulazione classica degli anni 1950-1970, si fondava su una distinzione categoriale netta tra enunciati grammaticali e agrammaticali. Questa visione binaria presupponeva che la competenza linguistica di un parlante nativo potesse essere adeguatamente modellata mediante un sistema di regole formali che generassero l'insieme infinito delle frasi ben formate di una lingua, escludendo contestualmente tutte quelle mal formate. Il celebre esempio "Colorless green ideas sleep furiously", introdotto da Noam Chomsky nel 1957 all'interno di *Syntactic Structures*, incarnava paradigmaticamente questa posizione: pur essendo semanticamente anomalo, l'enunciato veniva considerato sintatticamente impeccabile, dimostrando così l'autonomia della sintassi rispetto al significato e la possibilità di tracciare una linea di demarcazione netta tra il grammaticale e l'agrammaticale.

Questo approccio si inseriva all'interno di un più ampio programma di ricerca che mirava a formalizzare la linguistica come scienza esatta, seguendo il modello delle scienze naturali e della matematica. La grammatica generativa, nel suo programma originario, si proponeva di specificare in modo esplicito e finito la conoscenza tacita che un parlante nativo possiede della propria lingua. Tale conoscenza veniva concepita come un sistema computazionale che associa rappresentazioni fonologiche a rappresentazioni semantiche attraverso la mediazione di strutture sintattiche. La sintassi occupava dunque una posizione centrale e autonoma nel modello complessivo della facoltà linguistica, governata da principi formali indipendenti da considerazioni semantiche o pragmatiche.

Il modello standard della teoria generativa, sviluppato principalmente tra il 1965 e il 1970, articolava la grammatica in tre componenti principali: una componente sintattica che generava strutture profonde mediante regole di riscrittura (phrase structure rules), una componente trasformazionale che derivava strutture superficiali dalle strutture profonde mediante trasformazioni obbligatorie e facoltative, e componenti interpretative (semantica e fonologica) che operavano rispettivamente sulle strutture profonde e superficiali. La componente sintattica era concepita come il cuore generativo del sistema, mentre semantica e fonologia avevano ruolo puramente interpretativo.

Le regole di riscrittura della componente sintattica avevano la forma generale A → B C, dove A rappresenta un simbolo non terminale (categoria sintattica) e B, C possono essere sia simboli terminali (lessemi) sia non terminali. Un semplice esempio di sistema di regole potrebbe includere:

S → NP VP VP → V NP NP → Det N Det → {the, a, this} N → {cat, dog, man} V → {saw, chased, loves}

Tale sistema generava frasi come "the cat saw a dog" ma escludeva stringhe come "*cat the saw dog a". La distinzione tra grammaticale e agrammaticale emergeva come conseguenza diretta dell'applicazione meccanica delle regole: una stringa era grammaticale se e solo se poteva essere generata dal sistema di regole, agrammaticale altrimenti. Non vi era spazio per gradazioni intermedie o zone grigie.

Tuttavia, già nelle prime fasi di sviluppo della teoria generativa emersero evidenze empiriche che problematizzavano questa concezione dicotomica. Gli studi condotti mediante giudizi di grammaticalità raccolti da parlanti nativi mostravano sistematicamente che numerosi fenomeni sintattici non si prestavano a una classificazione binaria, manifestando invece gradi variabili di accettabilità. Le costruzioni che coinvolgevano dipendenze a lunga distanza, come le relative incassate o le interrogative con isole sintattiche, mostravano pattern di gradienza che resistevano a qualsiasi tentativo di categorizzazione assoluta.

Un esempio paradigmatico è costituito dalle violazioni delle isole sintattiche: mentre alcune violazioni producono giudizi di inaccettabilità pressoché unanimi, altre generano risposte più sfumate, con variabilità sia inter-individuale che intra-individuale. Consideriamo le seguenti estrazioni da isole di tipo diverso:

(1) What did John wonder whether Mary bought? (*whether-island*) (2) Who did you meet the person who knows? (*complex NP island*) (3) How did Mary leave before John fixed the car? (*adjunct island*)

I giudizi di acceptability per questi esempi non si distribuiscono in modo bimodale (completamente accettabili vs completamente inaccettabili), ma piuttosto lungo un continuum, con (1) tipicamente giudicato il più severo, (3) il meno severo, e (2) in posizione intermedia. Inoltre, la severità della violazione può essere modulata da fattori aggiuntivi come la lunghezza della dipendenza, la presenza di materiale interveniente, e proprietà semantico-pragmatiche del contesto.

La distinzione operata da Chomsky stesso tra grammaticalità e acceptability rappresentava un primo tentativo di riconciliare la teoria formale con i dati empirici. La grammaticalità veniva concepita come una proprietà teorica, derivante dalla competenza linguistica idealizzata, mentre l'acceptability era vista come un fenomeno legato alla performance, influenzato da fattori extragrammaticali quali limitazioni di memoria di lavoro, preferenze di processamento, frequenza d'uso e convenzioni pragmatiche. Secondo questa distinzione, un enunciato poteva essere perfettamente grammaticale dal punto di vista della competenza ma inaccettabile nella performance a causa di limitazioni cognitive o comunicative.

Questa distinzione, tuttavia, si rivelava problematica nella pratica empirica: come stabilire dove termina la grammaticalità e dove inizia la performance quando gli stessi giudizi di

grammaticalità, strumento principale per accedere alla competenza linguistica, sono influenzati dai fattori di performance? Il celebre esempio di center-embedding multipli illustra il dilemma:

(4) The rat the cat the dog chased caught ate the cheese.

Tale frase è generata da regole grammaticali perfettamente legittime della grammatica inglese (relative restrittive applicate ricorsivamente), eppure risulta praticamente incomprensibile per i parlanti nativi. Dovremmo considerarla grammaticale (in quanto generata dalle regole) ma inaccettabile (a causa di limitazioni di processamento), oppure la sua inaccettabilità estrema riflette in qualche modo una violazione della competenza linguistica stessa? Se la competenza genera strutture che i parlanti non possono mai processare o produrre, in che senso tale competenza cattura la conoscenza linguistica reale?

Il Problema dei Giudizi di Grammaticalità e la Natura Graduale dell'Acceptability

Metodologie Sperimentali per la Raccolta di Giudizi di Acceptability

La metodologia dei giudizi di grammaticalità, fondamento empirico della linguistica teorica per oltre 6 decenni, ha rivelato progressivamente la propria complessità intrinseca. Le tecniche tradizionali si basavano su giudizi informali forniti da linguisti stessi o da piccoli gruppi di parlanti nativi, spesso in formato binario (grammaticale vs agrammaticale) oppure ternario (accettabile, marginale, inaccettabile, indicati rispettivamente con simboli ✓, ? e *). Tuttavia, questi metodi presentavano problemi significativi di affidabilità e replicabilità.

Quando ai parlanti nativi viene chiesto di valutare l'accettabilità di enunciati su scale Likert (tipicamente a 5 o 7 punti) oppure mediante magnitude estimation (una tecnica mutuata dalla psicofisica che consente valutazioni su scala continua), i dati raccolti mostrano distribuzioni continue piuttosto che bimodali. La tecnica di magnitude estimation, introdotta nella linguistica sperimentale da Bard, Robertson e Sorace nel 1996, merita particolare attenzione per le sue proprietà metodologiche.

In magnitude estimation, ai partecipanti viene presentato un enunciato di riferimento (anchor) con un valore numerico arbitrario (per esempio, 100), e viene chiesto di assegnare valori numerici ad altri enunciati in proporzione alla loro acceptability relativa rispetto all'anchor. Se un enunciato è giudicato due volte più accettabile dell'anchor, dovrebbe ricevere valore 200; se è giudicato la metà accettabile, valore 50. Questa tecnica consente ai partecipanti di esprimere differenze finemente graduate senza essere vincolati a categorie predefinite.

Studi sperimentali condotti da Bard, Robertson e Sorace (1996) utilizzando magnitude estimation hanno dimostrato che i parlanti producono giudizi finemente graduati che si distribuiscono lungo un continuum, con correlazioni significative tra giudicanti che suggeriscono l'esistenza di intuizioni condivise riguardo ai gradi di (in)accettabilità. Tipicamente, le correlazioni inter-soggetto misurate tramite coefficiente di correlazione di

Pearson raggiungono valori nell'ordine di r = 0,70 - 0,85, indicando sostanziale accordo pur con variazione individuale.

Un aspetto importante emerso da questi studi è che la variabilità inter-individuale non è casuale ma sistematica: alcuni parlanti tendono a essere più permissivi nel loro giudizio complessivo (assegnando valori più alti), mentre altri sono più restrittivi, ma il ranking relativo delle costruzioni tende a essere preservato. Questo può essere modellato matematicamente mediante trasformazioni lineari dei punteggi individuali. Se indichiamo con x_{ij} il giudizio del soggetto i sull'enunciato j, possiamo standardizzare mediante z-score:

z_{ij} = (x_{ij} - μ_i) / σ_i

dove μ_i è la media dei giudizi del soggetto i e σ_i la deviazione standard. Dopo questa trasformazione, le correlazioni tra soggetti tipicamente aumentano a r > 0,85, confermando che la variazione riflette principalmente differenze nella calibrazione della scala piuttosto che in genuine divergenze nelle intuizioni linguistiche.

Gradienza Categoriale e Gradienza Strutturale

La ricerca di Frank Keller (2000) sulla gradience sintattica ha sistematizzato queste osservazioni empiriche, distinguendo due tipi fondamentali di gradienza: la gradience categoriale e la gradience strutturale. Questa distinzione si rivela essenziale per comprendere i diversi modi in cui la gradualità si manifesta nella sintassi.

La gradience categoriale si riferisce alla possibilità che un elemento linguistico possegga proprietà di più categorie sintattiche in gradi diversi, sfumando i confini tra le categorie tradizionali. Un esempio classico è costituito dagli aggettivi che mostrano proprietà nominali in gradi variabili. In inglese, aggettivi possono essere nominalizzati mediante l'articolo determinativo:

(5a) The rich are getting richer. (nominalizzazione completa) (5b) The very rich are getting richer. (gradazione aggettivale mantenuta) (5c) *The very richness is increasing. (impossibile con nome puro)

Questi esempi mostrano che "the rich" occupa una posizione intermedia tra aggettivo puro e nome puro, mantenendo alcune proprietà di entrambe le categorie. Può essere modificato da avverbi di grado ("very") come gli aggettivi, ma funziona come argomento nominale del predicato. La sua categorizzazione non può essere ridotta a una scelta binaria aggettivo/nome ma richiede un modello che ammetta appartenenza graduale a categorie.

La gradience strutturale, invece, concerne la variazione nel grado di violazione di vincoli sintattici, dove alcune violazioni sono percepite come più gravi di altre. Keller ha dimostrato mediante esperimenti sistematici che molti vincoli sintattici tradizionalmente concepiti come categoriali producono effetti graduali quando violati. Per esempio, il vincolo contro l'estrazione dal soggetto (Subject Condition) mostra gradienza rispetto al tipo di elemento estratto e alla struttura dell'isola:

(6a) ?What did pictures of disturb John? (estrazione da PP in soggetto semplice) (6b) ??What did the pictures of disturb John? (con determinante) (6c) ???What did several pictures of disturb John the most? (con quantificatore)

I giudizi di acceptability decrescono progressivamente da (6a) a (6c), suggerendo che fattori come definitezza e quantificazione modulano la severità della violazione. Keller ha raccolto dati di magnitude estimation che confermano questa gradazione, con (6a) che riceve tipicamente punteggi attorno a 60-70 (normalizzando la frase completamente accettabile a 100), (6b) attorno a 40-50, e (6c) sotto 30.

Un esempio emblematico della gradience strutturale è rappresentato dalle violazioni delle isole sintattiche in contesti diversi. L'estrazione da una whether-island produce tipicamente giudizi più severi rispetto all'estrazione da alcune isole complesse come i relativi soggetto, che a loro volta sono peggiori dell'estrazione da relativi oggetto:

(7a) What did you wonder whether John bought? (whether-island, severità massima) (7b) What did you meet the person who bought? (complex NP island, severità intermedia) (7c) What did you meet the person that John bought for? (object relative, meno severo)

Studi sperimentali quantitativi hanno assegnato a queste costruzioni valori medi di acceptability (normalizzati) rispettivamente di circa 25, 45 e 65 su scala 0-100, dimostrando una chiara graduazione.

Fenomeni di Gradienza nelle Costruzioni a Controllo

I fenomeni di gradience sintattica sono stati documentati estensivamente in molteplici domini empirici. Le costruzioni a controllo, per esempio, mostrano una scala continua di preferenze che dipende dalla semantica del verbo matrice. Il fenomeno del controllo riguarda l'identificazione del soggetto implicito di un verbo non finito in frasi come:

(8) John persuaded Mary to leave.

dove il soggetto di "leave" è controllato da "Mary" (il complemento oggetto di "persuade"). Diversi verbi mostrano preferenze di forza variabile per il controllo del soggetto versus dell'oggetto. Verbi come "force" impongono il controllo dell'oggetto in modo estremamente stringente:

(9a) John forced Mary to leave. (Mary lascia, controllo oggetto) (9b) ??John forced Mary for him to leave. (tentativo di evitare controllo, fortemente degradato)

Altri verbi come "persuade" mostrano preferenza forte ma non assoluta per il controllo dell'oggetto:

(10a) John persuaded Mary to leave. (preferito, Mary lascia) (10b) ?John persuaded Mary that he should leave. (possibile ma meno preferito)

Verbi come "want" permettono più liberamente alternative strutturali:

(11a) John wants to leave. (controllo soggetto) (11b) John wants Mary to leave. (controllo oggetto) (11c) John wants for Mary to leave. (forma esplicita)

Questo pattern può essere formalizzato assegnando a ciascun verbo un peso probabilistico che riflette la forza con cui impone il controllo. Se indichiamo con P(control=object|verb) la probabilità che un dato verbo imponga il controllo dell'oggetto quando presente, possiamo osservare empiricamente distribuzioni come:

P(control=object|force) ≈ 0,98 P(control=object|persuade) ≈ 0,85 P(control=object|want) ≈ 0,45

Questi valori sono derivati da analisi di corpora annotati e da studi di priming sintattico. Bresnan e colleghi hanno raccolto dati da corpus contenenti oltre 50.000 istanze di costruzioni infinitivali e hanno calcolato le frequenze relative delle diverse realizzazioni per ciascun verbo. Le probabilità empiriche mostrano distribuzione continua piuttosto che categorizzazione binaria in verbi di controllo oggetto vs soggetto.

Inoltre, studi di priming sintattico dimostrano che l'esposizione a costruzioni di controllo con un particolare verbo aumenta la probabilità di riutilizzare la stessa struttura con lo stesso verbo, e l'entità del priming è proporzionale alla "sorpresa" della struttura (inversamente proporzionale alla sua probabilità baseline). Se indichiamo con P_0(structure|verb) la probabilità baseline di una struttura con un verbo, e con P_primed(structure|verb) la probabilità dopo priming, gli studi mostrano tipicamente:

P_primed(structure|verb) = P_0(structure|verb) + α × (1 - P_0(structure|verb))

dove α ≈ 0,15-0,25 è il coefficiente di priming. Strutture rare (basso P_0) mostrano priming maggiore in termini assoluti.

Dipendenze Filler-Gap e Modularità delle Violazioni di Isola

Un altro dominio che esemplifica la natura graduale della sintassi è costituito dalle dipendenze filler-gap, ovvero costruzioni in cui un elemento (il filler) appare in una posizione diversa dalla sua posizione canonica, lasciando un "gap" nella struttura superficiale. Esempi includono interrogative wh-, relativizzazioni e topicalizzazioni:

(12a) What did John buy _? (interrogativa wh-) (12b) the book that Mary read _ (relativizzazione) (12c) This book, I really enjoyed _ (topicalizzazione)

La teoria delle isole sintattiche, formulata originariamente da John Ross nel 1967 nel suo lavoro seminale "Constraints on Variables in Syntax" e raffinata da Chomsky nella teoria delle barriere, postulava vincoli categoriali sull'estrazione. Ross identificò numerose configurazioni strutturali da cui l'estrazione risulta impossibile o fortemente degradata, tra cui:

- Complex NP Constraint: impossibile estrarre da nominali complessi
- Coordinate Structure Constraint: impossibile estrarre da un singolo congiunto
- Sentential Subject Constraint: impossibile estrarre da soggetti frasali
- Wh-Island Constraint: impossibile estrarre attraverso complementi interrogativi

Nella formulazione originale, questi vincoli erano concepiti come condizioni assolute: una struttura che violava un vincolo era categorialmente esclusa dalla grammatica. Tuttavia,

studi sperimentali successivi hanno ripetutamente dimostrato che le violazioni delle isole producono effetti graduali piuttosto che categoriali.

Hofmeister e Sag (2010) hanno condotto esperimenti sistematici di magnitude estimation testando estrazioni da diverse configurazioni isola, manipolando simultaneamente la lunghezza della dipendenza filler-gap (misurata in numero di parole o di clause intervenute) e la complessità strutturale del materiale interveniente. I loro risultati rivelano pattern di additività nelle penalità di acceptability: la riduzione di acceptability dovuta alla violazione di un'isola si somma approssimativamente in modo lineare alla riduzione dovuta alla lunghezza della dipendenza.

Formalmente, se modelliamo l'acceptability come funzione di tipo isola, distanza e complessità, otteniamo una relazione approssimativamente additiva in scala logaritmica:

log(acceptability) ≈ β_0 - β_1 × island_severity - β_2 × log(distance) - β_3 × complexity

dove i coefficienti β vengono stimati mediante regressione lineare sui dati sperimentali. Gli studi di Hofmeister e Sag hanno trovato valori come:

β_0 ≈ 4,5 (intercetta, acceptability baseline) β_1 ≈ 0,8-1,5 (penalità per tipo isola, varia tra tipi) β_2 ≈ 0,3-0,5 (penalità per distanza) β_3 ≈ 0,2-0,4 (penalità per complessità)

Questo pattern additivo suggerisce che le isole non costituiscono vincoli categoriali hard ma piuttosto penalità graduali che si compongono con altri fattori di processamento. Inoltre, la penalità β_1 varia sistematicamente tra tipi di isola diversi. Una scala approssimativa di severità emersa dagli studi è:

whether-islands: β_1 ≈ 1,4 complex NP islands (subject): β_1 ≈ 1,2 adjunct islands: β_1 ≈ 0,9 complex NP islands (object): β_1 ≈ 0,7

Questi risultati hanno implicazioni teoriche profonde. Se le isole sintattiche costituissero vincoli categoriali della competenza grammaticale, non dovremmo osservare gradazioni sistematiche né additività con fattori di performance. Il fatto che la severità delle violazioni vari continuamente e si componga con fattori di processamento suggerisce piuttosto che i fenomeni isola emergano dall'interazione di vincoli multipli, alcuni potenzialmente grammaticali e altri legati all'architettura del sistema di processamento.

Modelli Probabilistici della Competenza Linguistica: Dal Simbolico allo Stocastico

Transizione Paradigmatica: Dalla Categoricità alla Probabilità

L'emergere di evidenze empiriche a favore della gradienza sintattica ha motivato lo sviluppo di modelli formali che incorporano la probabilità come componente fondamentale della rappresentazione grammaticale. A differenza dell'approccio generativista classico, che trattava la probabilità come epifenomeno della performance (mero riflesso di fattori extragrammaticali che oscurano una competenza categoriale sottostante), i modelli

probabilistici considerano le distribuzioni di probabilità come parte integrante della competenza linguistica stessa.

Questa transizione rappresenta uno shift paradigmatico fondamentale nel modo di concepire la conoscenza linguistica. Invece di chiedersi "questa stringa è grammaticale?", la domanda diventa "qual è la probabilità di questa stringa data la grammatica?". Questa riformulazione non è meramente terminologica ma riflette una concezione radicalmente diversa della natura della competenza linguistica.

Nel paradigma categoriale, la grammatica specifica un insieme discreto di strutture ben formate, escludendo tutte le altre. La competenza è modellata come funzione caratteristica di tale insieme:

g(s) = 1 se s ∈ L (linguaggio generato dalla grammatica) g(s) = 0 se s ∉ L

Nel paradigma probabilistico, la grammatica specifica una distribuzione di probabilità sullo spazio di tutte le possibili stringhe:

G: Σ* → [0,1] tale che ∑_{s ∈ Σ*} G(s) = 1

dove Σ* denota l'insieme di tutte le stringhe finite sull'alfabeto Σ. La competenza non è più una funzione binaria ma una funzione continua che assegna probabilità diverse a strutture diverse. Stringhe altamente probabili corrispondono a costruzioni centrali, prototipiche della lingua, mentre stringhe a bassa probabilità sono periferiche o marginali. Il confine tra grammaticale e agrammaticale diventa sfumato, sostituito da un continuum di (in)acceptability.

Modelli N-Grammi: Fondamenti e Limitazioni

I modelli probabilistici più elementari della sintassi sono costituiti dai modelli n-grammi, che stimano la probabilità di una sequenza di parole mediante approssimazioni markoviane. L'idea fondamentale è sfruttare la decomposizione della probabilità congiunta mediante la regola della catena:

P(w_1, w_2, ..., w_n) = P(w_1) × P(w_2|w_1) × P(w_3|w_1,w_2) × ... × P(w_n|w_1,...,w_{n-1})

Il calcolo esatto di questa espressione richiede la stima di P(w_i|w_1,...,w_{i-1}) per ogni possibile storia di lunghezza variabile, il che è computazionalmente intrattabile e statisticamente impossibile dato che la maggior parte delle storie lunghe non occorre mai in nessun corpus per quanto grande.

I modelli n-grammi risolvono questo problema mediante l'assunzione markoviana di ordine n-1: la probabilità della parola corrente dipende solo dalle n-1 parole precedenti, non dall'intera storia:

P(w_i|w_1,...,w_{i-1}) ≈ P(w_i|w_{i-n+1},...,w_{i-1})

Un modello bigramma (n=2) assume quindi:

$P(w_1, w_2, ..., w_m) \approx P(w_1) \times P(w_2|w_1) \times P(w_3|w_2) \times ... \times P(w_m|w_{m-1})$

Applicando il logaritmo per ragioni di stabilità numerica (i prodotti di molte probabilità piccole causano underflow), si ottiene:

$\log P(w_1, w_2, ..., w_m) = \log P(w_1) + \sum_{i=2}^{m} \log P(w_i|w_{i-1})$

I parametri di tali modelli, ovvero le probabilità condizionali $P(w_i|w_{i-1})$, possono essere stimate mediante massima verosimiglianza da corpora annotati. Dato un corpus C, la stima di massima verosimiglianza è semplicemente la frequenza relativa:

$\hat{P}(w_i|w_{i-1}) = \text{count}(w_{i-1}, w_i) / \text{count}(w_{i-1})$

dove $\text{count}(w_{i-1}, w_i)$ indica il numero di occorrenze della sequenza w_{i-1} w_i nel corpus e $\text{count}(w_{i-1})$ il numero di occorrenze totali di w_{i-1}.

Per esempio, dato un corpus che contiene 1.000.000 di parole, di cui "the" occorre 50.000 volte e la sequenza "the cat" occorre 500 volte, stimiamo:

$\hat{P}(\text{cat}|\text{the}) = 500 / 50.000 = 0,01$

Nonostante la loro semplicità computazionale, i modelli n-grammi catturano alcuni aspetti della gradienza sintattica. Per esempio, la sequenza "the cat sat on the mat" avrà probabilità significativamente più alta rispetto a "mat the on sat cat the" anche in un semplice modello bigramma, in virtù delle probabilità condizionali apprese dal corpus. Calcoliamo le probabilità utilizzando stime ipotetiche da un corpus inglese:

Per "the cat sat on the mat": $\log P \approx \log(0,05) + \log(0,01) + \log(0,02) + \log(0,15) + \log(0,08) + \log(0,001) \approx -3,0 - 4,6 - 3,9 - 1,9 - 2,5 - 6,9 = -22,8$

Per "mat the on sat cat the": $\log P \approx \log(0,0001) + \log(0,00001) + \log(0,0001) + \log(0,00001) + \log(0,00001) + \log(0,08) \approx -9,2 - 11,5 - 9,2 - 11,5 - 11,5 - 2,5 = -55,4$

La differenza di circa 32,6 in log-probabilità corrisponde a un fattore di circa $\exp(32,6) \approx 1,4 \times 10^{14}$ nella probabilità, una differenza schiacciante che cattura l'intuizione che la prima frase è naturale mentre la seconda è completamente degradata.

Tuttavia, i modelli n-grammi presentano limiti intrinseci fondamentali:

1. **Non catturano dipendenze a lunga distanza**: Un modello bigramma non può catturare la dipendenza tra soggetto e verbo in frasi come "The cat that the dog chased (verb)" dove possono intervenire molte parole tra soggetto e verbo. Anche aumentando n (trigrammi, 4-grammi, etc.), le dipendenze che superano la finestra di contesto rimangono invisibili.

2. **Ignorano la struttura gerarchica**: I modelli n-grammi trattano il linguaggio come sequenza lineare, ignorando che le frasi hanno struttura gerarchica ad albero. La frase "The cat [that the dog chased] ran away" ha una struttura in cui "that the dog chased" è un costituente che modifica "cat", non semplicemente una sequenza di

parole adiacenti.

3. **Soffrono di data sparsity**: Per n elevati, il numero di possibili n-grammi cresce esponenzialmente (|V|^n dove V è il vocabolario), ma la maggior parte non occorrerà mai in nessun corpus. Con un vocabolario di 50.000 parole, vi sono 50.000^3 = 125 trilioni di possibili trigrammi, ma anche un corpus enorme come Google Books (circa 1 trilione di token) copre solo una minuscola frazione di questi.

4. **Assegnano probabilità zero a sequenze non viste**: Se un n-gramma particolare non occorre nel corpus di addestramento, la stima di massima verosimiglianza assegna probabilità zero. Questo è chiaramente inadeguato, dato che sequenze perfettamente grammaticali e naturali potrebbero non occorrere per pura casualità in un corpus finito.

Il problema della data sparsity viene parzialmente affrontato mediante tecniche di smoothing, che redistribuiscono massa di probabilità da n-grammi frequenti a n-grammi rari o non visti. Diverse tecniche sono state proposte:

Add-k smoothing (Laplace smoothing): Si aggiunge una pseudo-count k a tutti i conteggi:

$\hat{P}_{smooth}(w_i|w_{i-1}) = (count(w_{i-1}, w_i) + k) / (count(w_{i-1}) + k|V|)$

dove |V| è la dimensione del vocabolario. Tipicamente si usa k=1 (Laplace smoothing) o k<1.

Good-Turing smoothing: Utilizza la frequenza di frequenze (quanti n-grammi occorrono esattamente r volte) per stimare la probabilità di n-grammi non visti. L'idea è che se molti n-grammi occorrono esattamente una volta, vi sono probabilmente molti n-grammi con occorrenza zero che potrebbero occorrere una volta in più dati.

Kneser-Ney smoothing: Uno dei metodi più efficaci, che considera non solo la frequenza assoluta di un n-gramma ma anche il numero di contesti diversi in cui appare. Un bigramma "Francisco" può avere alta frequenza ma appare quasi sempre dopo "San", quindi non dovrebbe ricevere alta probabilità in altri contesti.

Nonostante questi miglioramenti, i limiti fondamentali dei modelli n-grammi rimangono. Per modellare adeguatamente la sintassi è necessario catturare la struttura gerarchica e le dipendenze a lunga distanza, il che richiede modelli più sofisticati.

Probabilistic Context-Free Grammars (PCFG): Struttura Gerarchica e Probabilità

Le Probabilistic Context-Free Grammars (PCFG) rappresentano un'estensione probabilistica delle grammatiche context-free (CFG) tradizionali che supera parzialmente i limiti dei modelli n-grammi incorporando struttura gerarchica. Una CFG è definita formalmente come una quadrupla G = (N, Σ, R, S) dove:

- N è un insieme finito di simboli non terminali (categorie sintattiche: S, NP, VP, etc.)

- Σ è un alfabeto finito di simboli terminali (parole del lessico)
- $R \subseteq N \times (N \cup \Sigma)^*$ è un insieme finito di regole di produzione
- $S \in N$ è il simbolo iniziale (tipicamente S per sentence)

Le regole di produzione hanno la forma $A \rightarrow \alpha$ dove $A \in N$ e $\alpha \in (N \cup \Sigma)^*$. Una CFG genera stringhe mediante derivazioni: partendo dal simbolo iniziale S, si applicano iterativamente regole di produzione sostituendo simboli non terminali finché non rimangono solo terminali.

Una PCFG estende questo formalismo associando a ciascuna regola di produzione una probabilità. Formalmente, una PCFG è una quintupla $G = (N, \Sigma, R, S, P)$ dove i primi quattro elementi sono come nella CFG e $P: R \rightarrow [0,1]$ è una funzione che assegna a ciascuna regola una probabilità, soggetta al vincolo di normalizzazione: per ogni simbolo non terminale A, le probabilità delle regole che lo espandono devono sommare a 1:

$$\sum_{(A \rightarrow \alpha) \in R} P(A \rightarrow \alpha) = 1$$

Questo vincolo assicura che le probabilità definiscano una distribuzione di probabilità propria sugli alberi sintattici.

La probabilità di un albero sintattico T è data dal prodotto delle probabilità di tutte le regole utilizzate nella sua derivazione. Se T utilizza le regole $r_1, r_2, ..., r_k$, allora:

$$P(T) = \prod_{i=1}^{k} P(r_i) = \prod_{(A \rightarrow \alpha) \in T} P(A \rightarrow \alpha)$$

La probabilità di una stringa w è ottenuta sommando (marginalizzando) le probabilità di tutti gli alberi sintattici possibili che la generano:

$$P(w) = \sum_{T: \text{yield}(T) = w} P(T)$$

dove yield(T) indica la stringa terminale (sequenza di foglie) prodotta dall'albero T. Questa somma può contenere infiniti termini se la grammatica consente ricorsione illimitata, ma tipicamente solo un numero finito di alberi contribuisce massa significativa alla probabilità totale.

I parametri di una PCFG, ovvero le probabilità $P(A \rightarrow \alpha)$ per ogni regola, possono essere stimati da treebanks (corpora di frasi annotate con struttura sintattica) mediante conteggio delle frequenze relative:

$$\hat{P}(A \rightarrow \alpha) = \text{count}(A \rightarrow \alpha) / \sum_{(A \rightarrow \beta) \in R} \text{count}(A \rightarrow \beta)$$

dove $\text{count}(A \rightarrow \alpha)$ è il numero di volte che la regola $A \rightarrow \alpha$ è utilizzata negli alberi del treebank. Questa è la stima di massima verosimiglianza dei parametri data l'osservazione degli alberi completi.

Le PCFG costituiscono un framework naturale per modellare la gradienza sintattica: strutture più probabili secondo la grammatica dovrebbero corrispondere a costruzioni più accettabili per i parlanti. L'intuizione è che la competenza linguistica include non solo conoscenza delle strutture possibili, ma anche delle loro probabilità relative, riflettendo l'esperienza accumulata con il linguaggio.

Uno studio seminale di Bresnan e collaboratori (2007) ha testato esplicitamente questa ipotesi nel contesto dell'alternanza dativa in inglese. L'alternanza dativa riguarda la possibilità di esprimere relazioni ditransitive in due modi:

(13a) John gave Mary a book. (double object dative) (13b) John gave a book to Mary. (prepositional dative)

Entrambe le costruzioni sono grammaticali, ma la loro acceptability relativa dipende da molteplici fattori inclusi animatezza, definitezza, pronominalità e peso sintattico degli argomenti. Bresnan et al. hanno addestrato una PCFG sul British National Corpus e hanno estratto le probabilità predette per costruzioni dative con varie combinazioni di proprietà degli argomenti. Hanno quindi raccolto giudizi di acceptability da parlanti nativi per le stesse costruzioni.

Il coefficiente di correlazione di Pearson tra le log-probabilità predette dal modello e i giudizi medi di acceptability era $r = 0{,}87$ ($p < 0{,}001$), indicando che le probabilità apprese dalla PCFG catturano effettivamente una porzione sostanziale della variazione nei giudizi di acceptability. Questo risultato fornisce evidenza forte che le distribuzioni probabilistiche riflettono aspetti della competenza linguistica, non solo della performance.

Tuttavia, le PCFG nella loro formulazione standard presentano assunzioni di indipendenza contestuale che limitano il loro potere espressivo e la loro adeguatezza cognitiva. In particolare, l'assunzione centrale è che la probabilità di espandere un non terminale dipenda solo dal suo simbolo categoriale, non dal contesto sintattico in cui si trova:

$$P(A \rightarrow \alpha \mid \text{context}) = P(A \rightarrow \alpha)$$

Questa assunzione è empiricamente inadeguata. Per esempio, la probabilità di espandere NP come pronome versus sintagma nominale complesso dipende fortemente dalla posizione sintattica: i pronomi sono molto più probabili in posizione soggetto che in posizione oggetto. In inglese scritto formale, dati approssimativi da corpus annotati mostrano:

$P(\text{NP} \rightarrow \text{pronoun} \mid \text{NP è soggetto}) \approx 0{,}45$ $P(\text{NP} \rightarrow \text{pronoun} \mid \text{NP è oggetto}) \approx 0{,}25$

Ma una PCFG standard non può catturare questa dipendenza perché non distingue NP in posizioni diverse: vi è un unico simbolo NP con un'unica distribuzione di espansioni.

Similmente, la probabilità di espansioni di VP dipende dalle proprietà del verbo testa (alcuni verbi richiedono complementi obbligatori, altri ammettono argomenti opzionali), ma una PCFG non lessicalizzata tratta tutti i VP allo stesso modo. Questi limiti hanno motivato lo sviluppo di raffinamenti che aumentano il potere espressivo delle PCFG.

Raffinamenti delle PCFG: Modelli Lessicalizzati e Parent Annotation

PCFG Lessicalizzate

Le Lexicalized PCFG, introdotte indipendentemente da Eugene Charniak (1997) e Michael Collins (1997) nel contesto del parsing statistico, incorporano informazione lessicale nelle

regole grammaticali, permettendo alla probabilità delle produzioni di dipendere dalle teste lessicali dei costituenti. L'idea centrale è che molto del comportamento sintattico è idiosincratico ai lessemi specifici, e catturare questa specificità lessicale migliora drasticamente le predizioni del modello.

In una PCFG lessicalizzata, ciascun non terminale è annotato con la sua testa lessicale h, e le regole hanno la forma:

$A(h) \rightarrow B(h)\ C(h_c)$

dove h indica la testa del costituente A (ereditata dal figlio testa, tipicamente B in una regola binaria head-initial), e h_c indica la testa del costituente complemento C. La probabilità di tale regola dipende sia dalle categorie sintattiche sia dalle teste lessicali:

$P(A(h) \rightarrow B(h)\ C(h_c) \mid A, h)$

oppure, scomponendo ulteriormente mediante assunzioni di indipendenza condizionale:

$P(A \rightarrow B\ C \mid A, h) \times P(h_c \mid A \rightarrow B\ C, h)$

La prima componente cattura la preferenza del verbo h per certe strutture sintattiche (subcategorizzazione), mentre la seconda cattura preferenze per specifici complementi lessicali (selezional restrictions).

Per esempio, il verbo "eat" seleziona con alta probabilità un complemento NP diretto (struttura transitiva), mentre preferisce meno strutture intransitive o con complementi frasali:

$P(VP \rightarrow V\ NP \mid VP, eat) \approx 0{,}75\ P(VP \rightarrow V \mid VP, eat) \approx 0{,}20\ P(VP \rightarrow V\ CP \mid VP, eat) \approx 0{,}05$

Al contrario, il verbo "think" preferisce complementi frasali:

$P(VP \rightarrow V\ CP \mid VP, think) \approx 0{,}65\ P(VP \rightarrow V \mid VP, think) \approx 0{,}30\ P(VP \rightarrow V\ NP \mid VP, think) \approx 0{,}05$

Inoltre, il modello può catturare preferenze per complementi lessicali specifici. Per esempio, "eat" co-occorre frequentemente con "food", "meal", "dinner", etc.:

$P(h_c = food \mid VP(eat) \rightarrow V\ NP, eat) > P(h_c = idea \mid VP(eat) \rightarrow V\ NP, eat)$

mentre per "think" la distribuzione su complementi nominali possibili è diversa:

$P(h_c = idea \mid VP(think) \rightarrow V\ NP, think) > P(h_c = food \mid VP(think) \rightarrow V\ NP, think)$

La lessicalizzazione cattura anche fenomeni di attachment ambiguity. Nella frase classicamente ambigua:

(14) I saw the man with the telescope.

il PP "with the telescope" può attaccarsi sia al VP (indicando lo strumento della visione) sia all'NP "the man" (indicando possesso). Una PCFG non lessicalizzata assegna probabilità

uniformi basate solo sulle categorie sintattiche. Una PCFG lessicalizzata può invece sfruttare le teste:

P(VP(saw) → VP(saw) PP(telescope)) vs P(NP(man) → NP(man) PP(telescope))

e inoltre considerare la plausibilità semantica delle combinazioni testa-modificatore, anche se quest'ultima è catturata solo indirettamente tramite frequenze di co-occorrenza nel corpus.

Il modello di Collins (1997), uno dei primi parser statistici ad alta performance, utilizza una decomposizione generativa complessa che stima sequenzialmente la probabilità di generare figli sinistro e destro della testa:

P(regola | A, h) = P(L_1, ..., L_m, R_1, ..., R_n, STOP | A, h)

dove L_i sono figli a sinistra della testa e R_i figli a destra, e STOP è un simbolo speciale che indica la terminazione della generazione di dipendenti. Questa probabilità viene ulteriormente scomposta assumendo che ciascun dipendente sia generato indipendentemente dagli altri dato il genitore:

P = ∏*{i=1}^{m} P(L_i(h_i) | A, h, distanza)* × *P(STOP_L | A, h)* × ∏{j=1}^{n} P(R_j(h_j) | A, h, distanza) × P(STOP_R | A, h)

dove distanza cattura la posizione relativa del dipendente (adiacente alla testa, una posizione rimossa, etc.). Questo modello ha ottenuto performance di parsing superiori a qualsiasi sistema precedente quando introdotto, raggiungendo F1-score di circa 88% sul Penn Treebank (section 23, il test set standard).

Parent Annotation e State Splitting

Un'altra tecnica di raffinamento, ortogonale alla lessicalizzazione, è la **parent annotation** proposta da Mark Johnson (1998). L'idea è annotare ciascun simbolo non terminale con il simbolo del suo genitore nell'albero sintattico, permettendo così alla grammatica di distinguere lo stesso non terminale in contesti sintattici diversi.

Una regola parent-annotated ha la forma:

A^P → B^A C^A

dove il superscript indica il genitore del non terminale. Per esempio, invece di un unico simbolo NP, si hanno NP^S (NP figlio di S, tipicamente soggetto), NP^VP (NP figlio di VP, tipicamente oggetto), NP^PP (NP figlio di PP, oggetto di preposizione), etc.

Questo raffinamento cattura differenze distribuzionali sistematiche. In inglese, NP in posizione soggetto (NP^S) hanno distribuzione diversa da NP in posizione oggetto (NP^VP):

- NP^S tendono più spesso a essere pronominali o nomi propri
- NP^S tendono a essere definiti
- NP^S sono in media più corti (meno modificatori)

Annotando il genitore, la PCFG può apprendere distribuzioni separate:

P(NP^S → pronoun) ≈ 0,45 P(NP^VP → pronoun) ≈ 0,25

P(NP^S → Det N) ≈ 0,35 P(NP^VP → Det N) ≈ 0,50

Klein e Manning (2003) hanno dimostrato sperimentalmente che l'aggiunta di parent annotation a una PCFG semplice aumenta significativamente la performance di parsing sul Penn Treebank. Partendo da una PCFG baseline con F1-score di 72,6%, l'aggiunta di parent annotation porta a 77,8%, un miglioramento sostanziale di 5,2 punti percentuali.

Inoltre, Klein e Manning hanno esplorato **horizontal markovization**, ovvero condizionare le espansioni sulla storia recente di fratelli già generati. Invece di generare tutti i figli indipendentemente, si mantiene memoria di alcuni figli precedenti:

P(regola) = P(A → α | A, genitore, fratelli_precedenti)

La combinazione di parent annotation verticale e markovization orizzontale porta ulteriori miglioramenti, raggiungendo F1 ≈ 80%.

Un ulteriore raffinamento consiste nell'utilizzare tecniche di **state splitting automatico**, come implementato nel Berkeley Parser di Petrov, Barrett, Thibaux e Klein (2006). Invece di specificare manualmente quali distinzioni (parent, grandparent, lessicalizzazione) sono rilevanti, l'algoritmo apprende automaticamente raffinamenti ottimali dei simboli non terminali mediante un processo iterativo di splitting e merging.

L'algoritmo procede come segue:

1. **Inizializzazione**: Si parte con una PCFG semplice (simboli base senza annotazioni)

2. **Split**: Ciascun simbolo non terminale A viene split in due versioni A_1, A_2. Le regole che coinvolgono A vengono duplicate per A_1 e A_2, e i parametri vengono inizializzati con piccole perturbazioni random attorno ai valori originali.

3. **EM**: Si esegue l'algoritmo Expectation-Maximization per stimare i parametri della grammatica split, massimizzando la verosimiglianza dei dati di training.

4. **Merge**: Si valuta per ciascuna coppia (A_1, A_2) la perdita di verosimiglianza che risulterebbe dal merge:

   ΔL(A_1, A_2) = log P(data | G_{merged}) - log P(data | G_{split})

   Se ΔL(A_1, A_2) > -ε per qualche soglia ε (ovvero la perdita è piccola), lo split viene annullato. Tipicamente ε è scelto per mantenere circa il 50% degli split.

5. **Iterazione**: Si ripete il processo di split-EM-merge per un numero prefissato di iterazioni (tipicamente 5-6).

Questo processo produce automaticamente raffinamenti che catturano generalizzazioni linguistiche importanti scoperte empiricamente. Per esempio, il simbolo NP viene tipicamente split in versioni che corrispondono approssimativamente a:

- NP_pronominal: pronomi personali (I, you, he, she, ...)
- NP_definite: sintagmi nominali con determinante definito (the cat, the government, ...)
- NP_indefinite: sintagmi nominali con determinante indefinito (a book, some people, ...)
- NP_bare: nomi nudi senza determinante (water, information, John, ...)
- NP_quantified: sintagmi nominali con quantificatori (many people, few students, ...)

Questi split emergono automaticamente dall'algoritmo perché catturano regolarità distribuzionali nei dati: i diversi sottotipi di NP hanno distribuzioni diverse di modificatori, compaiono in posizioni sintattiche diverse con frequenze diverse, e si combinano con verbi diversi.

Il Berkeley Parser con 5 cicli di split-merge raggiunge F1-score di circa 90% sul Penn Treebank, rappresentando uno stato dell'arte per parser basati puramente su PCFG (senza features aggiuntive o componenti discriminative).

Modelli Discriminativi e Maximum Entropy

Un approccio fondamentalmente diverso alla modellazione probabilistica della sintassi è costituito dai **modelli discriminativi**, che invece di modellare la probabilità congiunta P(T, w) di albero e stringa (come fanno le PCFG generative), modellano direttamente la probabilità condizionale P(T|w) dell'albero data la stringa.

La distinzione generativo vs discriminativo è importante sia metodologicamente che praticamente:

Modelli generativi:

- Modellano la distribuzione congiunta $P(T, w) = P(T) \times P(w|T)$ oppure $P(w) \times P(T|w)$
- Possono generare campioni (alberi e stringhe)
- Richiedono modellazione esplicita di P(w), la distribuzione delle stringhe
- Vincoli di indipendenza condizionale più stretti (per rendere P(w) trattabile)

Modelli discriminativi:

- Modellano solo P(T|w), la distribuzione degli alberi data la stringa osservata
- Non possono generare stringhe (ma questo non è necessario per il parsing)
- Non richiedono modellazione di P(w), solo della distribuzione condizionale
- Possono incorporare features arbitrarie e interdipendenti senza dover modellare la loro distribuzione marginale

I modelli **log-lineari** o **maximum entropy** rappresentano il framework più diffuso per modelli discriminativi. In un modello log-lineare, la probabilità di un albero T data una stringa w è definita come:

$P(T|w; \lambda) = (1 / Z(w; \lambda)) \times \exp(\sum_{i=1}^{k} \lambda_i f_i(T, w))$

dove:

- $f_i(T, w)$ sono funzioni feature che estraggono proprietà rilevanti dalla coppia albero-stringa
- $\lambda_i \in \mathbb{R}$ sono i pesi associati alle features (parametri da apprendere)
- $Z(w; \lambda)$ è la funzione di partizione che garantisce la normalizzazione:

  $Z(w; \lambda) = \sum_{T' \in Trees(w)} \exp(\sum_{i=1}^{k} \lambda_i f_i(T', w))$

  dove Trees(w) denota l'insieme di tutti i possibili alberi sintattici per la stringa w.

Le features f_i possono codificare informazioni arbitrariamente complesse. Alcuni esempi:

Features lessicali: $f_1(T, w) = 1$ se T contiene la regola $S \rightarrow NP\ VP$ e la testa di NP è "I", 0 altrimenti $f_2(T, w) = 1$ se T contiene $VP \rightarrow V\ NP$ con V = "eat" e testa(NP) = "food", 0 altrimenti

Features strutturali: $f_3(T, w)$ = numero di relative clause in T $f_4(T, w)$ = profondità massima dell'albero T $f_5(T, w)$ = numero di nodi unary (con un solo figlio) in T

Features di dipendenza: $f_6(T, w)$ = log(distanza tra filler e gap) se T contiene una dipendenza filler-gap, 0 altrimenti $f_7(T, w) = 1$ se T contiene estrazione da complex NP island, 0 altrimenti

Features di processamento: $f_8(T, w)$ = numero di center embeddings in T $f_9(T, w)$ = lunghezza della dipendenza soggetto-verbo più lunga in T

I modelli log-lineari offrono flessibilità estrema: si possono definire centinaia o migliaia di features catturando qualsiasi aspetto della struttura sintattica o della relazione albero-stringa che si ritenga rilevante. Non vi sono vincoli di indipendenza tra features: f_i e f_j possono essere arbitrariamente correlate.

I pesi λ vengono appresi mediante **massimizzazione della log-verosimiglianza condizionale** sul training set. Dato un training set $D = \{(w^{(j)}, T^{(j)})\}_{j=1}^{N}$ di coppie stringa-albero gold, l'obiettivo è trovare λ che massimizza:

$L(\lambda) = \sum_{j=1}^{N} \log P(T^{(j)}|w^{(j)}; \lambda) = \sum_{j=1}^{N} [\sum_{i=1}^{k} \lambda_i f_i(T^{(j)}, w^{(j)}) - \log Z(w^{(j)}; \lambda)]$

Il gradiente della log-verosimiglianza rispetto ai parametri è:

$\partial L/\partial \lambda_i = \sum_{j=1}^{N} [f_i(T^{(j)}, w^{(j)}) - \mathbb{E}_{T \sim P(T|w^{(j)};\lambda)} [f_i(T, w^{(j)})]]$

Il primo termine è la media empirica della feature i negli alberi gold del training set. Il secondo termine è il valore atteso della feature sotto la distribuzione del modello corrente. Il

gradiente è zero quando il modello assegna alle features le stesse medie che hanno nei dati gold - questa è la caratterizzazione del principio di maximum entropy.

L'ottimizzazione viene tipicamente effettuata mediante metodi del gradiente come L-BFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno), un metodo quasi-Newton che approssima la matrice Hessiana usando solo le informazioni dal gradiente. Si aggiunge tipicamente regolarizzazione per prevenire overfitting:

L_{reg}(λ) = L(λ) - (α/2) ||λ||_2^2

dove α > 0 è un iperparametro che controlla la forza della regolarizzazione L2 (ridge). Alternativamente si può usare regolarizzazione L1 per indurre sparsity (solo un sottoinsieme di features ha peso non zero).

Il calcolo della funzione di partizione Z(w; λ) e del valore atteso delle features sotto P(T|w; λ) è computazionalmente costoso perché richiede la somma su tutti i possibili alberi per w. Per una frase di lunghezza n, il numero di alberi sintattici binari è il numero di Catalan:

C_n = (1/(n+1)) × (2n choose n) ≈ 4^n / (n^{3/2} √π)

che cresce esponenzialmente. Fortunatamente, quando la grammatica sottostante è context-free, si possono utilizzare algoritmi di programmazione dinamica come l'**inside-outside algorithm** per calcolare efficientemente Z(w) e le marginal probabilities necessarie per i gradienti.

L'inside algorithm calcola ricorsivamente le probabilità inside β(i, j, A) = probabilità di generare le parole w_i...w_j dal non terminale A. Per una PCFG standard:

β(i, i+1, A) = P(A → w_i) (regole lessicali) β(i, j, A) = ∑*{A → BC} P(A → BC)* × ∑{k=i+1}^{j-1} β(i, k, B) × β(k, j, C) (regole binarie)

La probabilità totale della stringa è Z(w) = β(1, n, S). Per modelli log-lineari, si sostituisce P(A → BC) con exp(∑_i λ_i f_i) (features locali per quella regola) e si normalizza alla fine.

L'outside algorithm calcola le probabilità outside α(i, j, A) = probabilità di generare w_1...w_{i-1} A w_j...w_n dato S (ovvero il contesto esterno al costituente). Combinando inside e outside, si possono calcolare marginal probabilities di regole specifiche e quindi le expectation delle features.

La complessità computazionale è O(n^3 |G|) dove |G| è la dimensione della grammatica (numero di regole non terminali), la stessa complessità del CKY parsing algorithm. Questo è trattabile per frasi di lunghezza moderata (n ≤ 50-60 parole) su hardware moderno.

I modelli discriminativi hanno dimostrato performance superiori alle PCFG generative su molti task di parsing. Charniak e Johnson (2005) hanno introdotto un reranking model: prima si genera una lista di candidati parse usando una PCFG generativa (top-100), poi si rerank usando un modello discriminativo con features ricche. Questo approccio ha raggiunto F1 ≈ 91% sul Penn Treebank, significativamente sopra le pure PCFG.

Tuttavia, i modelli discriminativi hanno anche svantaggi:

- Computazionalmente più costosi (calcolo di Z(w) per ogni esempio di training, ripetuto a ogni iterazione)
- Possono soffrire di overfitting se il numero di features è molto grande rispetto ai dati
- Non modellano P(w), quindi non possono assegnare probabilità assolute a stringhe (solo rankin relativo di parse per una data stringa)

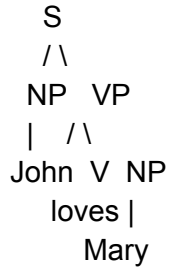
Data-Oriented Parsing e Modelli basati su Exemplar

Un approccio radicalmente diverso alla modellazione probabilistica della sintassi è rappresentato dal **Data-Oriented Parsing (DOP)**, introdotto da Rens Bod negli anni 1990. Il DOP rappresenta una rottura fondamentale con l'approccio generativista tradizionale basato su regole minimaliste e astratte. Invece, il DOP si fonda sull'idea che la competenza linguistica consista in un vasto inventario di frammenti arborali di dimensioni variabili estratti dall'esperienza linguistica. Questi frammenti (subtrees) possono essere molto piccoli (singole regole di produzione) o molto grandi (intere frasi idiomatiche o costruzioni multi-clausali), e la loro distribuzione riflette la frequenza di esposizione.

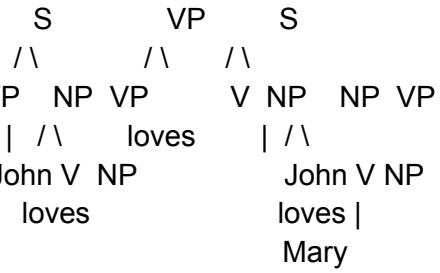
Formulazione Formale del DOP

Nel DOP, un corpus di alberi sintattici annotati C = {T_1, T_2, ..., T_N} viene analizzato per estrarre tutti i possibili subtree. Un subtree di un albero T è un frammento che include un nodo A e alcuni (non necessariamente tutti) dei suoi discendenti, dove i nodi non espansi diventano siti di sostituzione (indicati tipicamente con simboli come •).

Per esempio, dato l'albero per "John loves Mary":



alcuni possibili subtree includono:



Il numero totale di subtree in un albero di dimensione n è esponenziale: approssimativamente O(2^n) nel caso peggiore, anche se il numero di subtree distinti è generalmente inferiore grazie a ripetizioni strutturali.

La probabilità di un subtree t è definita come la sua frequenza relativa tra tutti i subtree con la stessa radice:

P(t) = count(t) / ∑_{t': root(t') = root(t)} count(t')

dove count(t) è il numero di occorrenze del subtree t negli alberi del corpus, e la somma al denominatore è su tutti i subtree con la stessa categoria radice di t.

Un albero sintattico T per una nuova stringa w può essere derivato (composto) mediante sequenze di operazioni di sostituzione che combinano subtree. Una derivazione d è una sequenza di subtree t_1, t_2, ..., t_k dove t_1 ha radice S (il simbolo iniziale), e ciascun t_{i+1} viene inserito in un sito di sostituzione di uno dei subtree precedenti. La derivazione termina quando non rimangono siti di sostituzione e le foglie sono tutte terminali che corrispondono a w.

Per un dato albero T, tipicamente esistono molteplici derivazioni diverse (potenzialmente infinite se si ammettono subtree di dimensione 1, che permettono ricombinazioni arbitrarie). La probabilità dell'albero T è la somma delle probabilità di tutte le derivazioni possibili:

P(T|C) = ∑_{d ∈ D(T,C)} P(d)

dove D(T,C) è l'insieme di tutte le derivazioni di T utilizzando subtree dal corpus C. La probabilità di una derivazione è il prodotto delle probabilità dei subtree utilizzati:

P(d) = P(t_1) × P(t_2) × ... × P(t_k) = ∏_{i=1}^{k} P(t_i)

Espandendo la definizione di P(t_i):

P(d) = ∏*{i=1}^{k} [count(t_i) /* ∑{t': root(t') = root(t_i)} count(t')]

Catturare la Gradienza tramite Compositività Probabilistica

Il DOP cattura naturalmente la gradienza sintattica attraverso un meccanismo elegante: costruzioni frequenti nel corpus possono essere composte utilizzando subtree di grandi dimensioni che occorrono frequentemente (chunks ad alta probabilità), risultando in probabilità complessiva alta. Costruzioni rare devono essere composte da subtree più piccoli e generali (che hanno probabilità individuale più bassa perché il denominatore include molti altri subtree concorrenti), risultando in probabilità complessiva inferiore.

Consideriamo un esempio concreto. La frase comune "I saw the cat" potrebbe essere presente molte volte nel corpus esattamente in questa forma. Esisterebbe quindi un grande subtree:

S
/ \
NP VP

```
|   / \
I  V  NP
   saw / \
      Det N
      the cat
```

con frequenza elevata, diciamo count = 150. Se il corpus contiene 1000 alberi con radice S totali, tutti i subtree con radice S sommano a (diciamo) 50.000 (contando molteplicità). Quindi:

P(questo_subtree_grande) = 150 / 50.000 = 0,003

Questo subtree può derivare l'intera frase in un singolo passo, quindi:

P("I saw the cat" derivata così) = 0,003

Ora consideriamo una variante rara, mai vista esattamente in questa forma nel corpus: "the cat saw I". Questa frase dovrà essere composta da subtree più piccoli. Per esempio:

1. Un subtree per la struttura base S → NP VP: P ≈ 0,8 (molto comune)
2. Un subtree per NP → Det N con Det=the, N=cat: P ≈ 0,002 (relativamente comune)
3. Un subtree per VP → V NP con V=saw: P ≈ 0,15 (comune)
4. Un subtree per NP → I: P ≈ 0,05 (pronome, comune ma non in oggetto)

La probabilità totale di questa derivazione è:

P ≈ 0,8 × 0,002 × 0,15 × 0,05 = 0,0000012

che è circa 2500 volte più bassa della frase naturale, catturando l'intuizione che questa costruzione è fortemente degradata.

Il meccanismo fondamentale è che subtree grandi e specifici (che catturano costruzioni idiomatiche o altamente frequenti) hanno probabilità individuale più alta rispetto a quanto si otterrebbe moltiplicando le probabilità dei loro componenti atomici. Questo riflette l'idea di **entrenchment**: costruzioni frequenti diventano unità consolidate (chunked) nella competenza linguistica, accessibili come interi piuttosto che composte part-by-part.

Sfide Computazionali e Approssimazioni

La formulazione DOP presenta però complessità computazionale proibitiva per l'implementazione diretta:

1. **Numero di subtree**: Il numero di subtree possibili estratti da un corpus anche modesto è astronomico. Un corpus di 10.000 frasi può facilmente produrre milioni o decine di milioni di subtree distinti.

2. **Numero di derivazioni**: Per un dato albero target T, il numero di derivazioni possibili può essere infinito. Se si ammettono subtree arbitrariamente piccoli (incluse singole regole), si possono creare derivazioni arbitrariamente lunghe che raggiungono lo

stesso albero T.

3. **Calcolo della somma**: Anche se limitiamo le derivazioni a lunghezza finita, la somma $\sum_{d \in D(T,C)} P(d)$ include potenzialmente miliardi di termini per alberi di dimensione moderata.

Bod ha proposto varie approssimazioni computazionalmente trattabili:

DOP1 (Bod 1992): Limita i subtree a profondità massima 2, riducendo drasticamente il numero di subtree. Questo però sacrifica la caratteristica distintiva del DOP (subtree di dimensione arbitraria).

Monte Carlo DOP: Stima P(T) mediante campionamento random di derivazioni. Si generano N derivazioni casuali per T usando un algoritmo di campionamento, e si approssima:

$P(T) \approx (1/N) \times \sum_{i=1}^{N} P(d_i)$

Con N sufficientemente grande (tipicamente 1000-10000), la stima converge al valore vero per la legge dei grandi numeri.

DOP-reduction: Converte il modello DOP in una PCFG compatta equivalente, eliminando la ridondanza nelle derivazioni. Tuttavia, questo richiede tecniche sofisticate per preservare le probabilità corrette.

Tree-DOP (Bod 2001): Utilizza una decomposizione gerarchica dell'insieme di subtree, organizzandoli in livelli di astrazione. Si campionano subtree a livelli diversi con probabilità dipendente dalla loro frequenza e generalità.

Nonostante queste approssimazioni, il DOP rimane computazionalmente costoso rispetto a PCFG standard, e per questo motivo non è stato ampiamente adottato nei sistemi di parsing pratici su larga scala. Tuttavia, le sue intuizioni fondamentali - che la competenza include frammenti di dimensione variabile con frequenze differenziate, e che la composizionalità è probabilistica - hanno influenzato profondamente lo sviluppo successivo di modelli linguistici.

Modelli Neurali e Deep Learning per la Sintassi

L'avvento del deep learning a partire dagli anni 2010 ha rivoluzionato la modellazione probabilistica del linguaggio, introducendo architetture neurali che apprendono rappresentazioni distribuite continue di parole, frasi e strutture sintattiche. I modelli neurali offrono vantaggi significativi rispetto agli approcci simbolici tradizionali:

1. **Gestione naturale della data sparsity**: Mediante generalizzazione basata su similarità in spazi vettoriali continui, i modelli neurali possono fare predizioni ragionevoli per combinazioni mai viste, interpolando tra esempi simili visti nel training.

2. **Apprendimento automatico di features**: Invece di specificare manualmente quali features sono rilevanti (come nei modelli log-lineari), le reti neurali apprendono automaticamente rappresentazioni hierarchiche di complessità crescente.

3. **Cattura di interazioni non lineari**: Le funzioni di attivazione non lineari (sigmoid, tanh, ReLU) permettono di modellare relazioni complesse e non lineari tra input e output.

4. **Scalabilità**: I moderni acceleratori hardware (GPU, TPU) rendono l'addestramento di reti neurali su dataset enormi computazionalmente fattibile.

Recurrent Neural Networks per Language Modeling

Le Recurrent Neural Networks (RNN) rappresentano la prima classe di modelli neurali applicati con successo alla modellazione del linguaggio. Una RNN processa una sequenza di input x_1, x_2, ..., x_n (dove x_t è tipicamente la rappresentazione vettoriale della parola al tempo t) mantenendo uno stato nascosto h_t che viene aggiornato ricorsivamente:

h_t = tanh(W_{hh} h_{t-1} + W_{xh} x_t + b_h)

dove:

- W_{hh} ∈ ℝ^{d×d} è la matrice di transizione stato-stato
- W_{xh} ∈ ℝ^{d×d_x} è la matrice input-stato
- b_h ∈ ℝ^d è il bias
- d è la dimensionalità dello stato nascosto
- d_x è la dimensionalità dell'input (embedding size)

L'output al tempo t, che rappresenta la distribuzione di probabilità sulla parola successiva, è calcolato come:

o_t = W_{ho} h_t + b_o y_t = softmax(o_t)

dove:

- W_{ho} ∈ ℝ^{|V|×d} mappa dallo stato nascosto allo spazio del vocabolario
- |V| è la dimensione del vocabolario
- softmax(z)_i = exp(z_i) / ∑_j exp(z_j) normalizza in distribuzione di probabilità

Per un language model, y_t rappresenta P(w_{t+1} | w_1, ..., w_t), la distribuzione sulla parola successiva dato il contesto precedente. La probabilità di una sequenza completa è:

P(w_1, w_2, ..., w_n) = ∏*{t=1}^{n} P(w_t | w_1, ..., w*{t-1}) = ∏*{t=1}^{n} y*{t-1}[w_t]

dove y_{t-1}[w_t] denota la componente di y_{t-1} corrispondente alla parola w_t.

L'addestramento avviene mediante massimizzazione della log-likelihood (equivalentemente, minimizzazione della cross-entropy):

L = -∑{t=1}^{n} log P(w_t | w_1, ..., w{t-1}) = -∑{t=1}^{n} log y{t-1}[w_t]

I gradienti sono calcolati mediante backpropagation through time (BPTT), che "srotola" la RNN nel tempo e applica backpropagation standard alla rete risultante.

Il Problema del Vanishing/Exploding Gradient

Le RNN vanilla soffrono del problema del vanishing/exploding gradient, che limita la loro capacità di catturare dipendenze a lunga distanza. Il gradiente di L rispetto ai parametri a tempo t dipende dal prodotto di Jacobiani:

∂L/∂W_{hh} = ∑{t=1}^{n} ∂L/∂y_t × (∂y_t/∂h_t) × (∏{k=1}^{t-1} ∂h_{k+1}/∂h_k) × ∂h_1/∂W_{hh}

Il termine critico è ∏{k=1}^{t-1} ∂h{k+1}/∂h_k. Per la funzione di aggiornamento h_t = tanh(W_{hh} h_{t-1} + ...), si ha:

∂h_{k+1}/∂h_k = diag(tanh'(W_{hh} h_k + W_{xh} x_{k+1} + b_h)) × W_{hh}

dove tanh'(z) = 1 - tanh^2(z) ∈ [0,1]. Il prodotto di molte matrici con elementi ≤ 1 tende esponenzialmente a zero (vanishing), mentre se alcuni autovalori di W_{hh} sono > 1, il prodotto può crescere esponenzialmente (exploding).

Questo significa che il gradiente per dipendenze che spans molti time steps (grande t) diventa o trascurabilmente piccolo (la rete non apprende le dipendenze lunghe) o esplode (instabilità numerica). Tipicamente, le RNN vanilla possono catturare dipendenze fino a 5-10 time steps ma falliscono per dipendenze più lunghe.

Long Short-Term Memory Networks (LSTM)

Le Long Short-Term Memory networks, introdotte da Hochreiter e Schmidhuber nel 1997, risolvono il problema del vanishing gradient mediante una architettura più complessa con gate di controllo che regolano il flusso di informazione. Un'unità LSTM mantiene due stati: il cell state C_t (memoria a lungo termine) e l'hidden state h_t (output). L'aggiornamento procede attraverso quattro passaggi:

1. Forget gate (decide quale informazione scartare dalla cell memory):

f_t = σ(W_f [h_{t-1}, x_t] + b_f)

dove σ è la funzione sigmoide σ(z) = 1/(1 + exp(-z)) ∈ [0,1], e [h_{t-1}, x_t] denota concatenazione dei vettori.

2. Input gate (decide quale nuova informazione memorizzare):

i_t = σ(W_i [h_{t-1}, x_t] + b_i) C̃_t = tanh(W_C [h_{t-1}, x_t] + b_C)

dove C̃_t è il candidate cell state.

3. Cell state update (combina forget e input):

$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$

dove $\odot$ denota prodotto elemento-per-elemento (Hadamard product). Questa è l'equazione chiave: quando $f_t \approx 1$ e $i_t \approx 0$, la cell memory è preservata intatta ($C_t \approx C_{t-1}$), permettendo propagazione del gradiente senza degradazione.

4. Output gate (decide cosa emettere):

$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o)$ $h_t = o_t \odot \tanh(C_t)$

L'idea chiave è che il cell state C_t ha un percorso diretto da C_{t-1} con solo moltiplicazione elemento-per-elemento per f_t. Se $f_t \approx 1$, il gradiente può fluire all'indietro senza decadimento esponenziale. Matematicamente:

$\partial C_t/\partial C_{t-1} = \text{diag}(f_t)$

che è ben condizionato (autovalori vicini a 1) quando i forget gate sono aperti. Questo permette alle LSTM di apprendere dipendenze su centinaia di time steps.

Le LSTM hanno dimostrato capacità notevoli nell'assegnare probabilità graduali a costruzioni sintattiche di complessità variabile. Studi hanno mostrato che LSTM language models assegnano perplexity (misura inversa di probabilità) più bassa a frasi sintatticamente ben formate rispetto a varianti agrammaticali, anche per fenomeni sintattici complessi come l'agreement a lunga distanza:

(15a) The keys to the cabinet are on the table. (corretto) (15b) *The keys to the cabinet is on the table. (violazione agreement)

LSTM language models assegnano log-probabilità significativamente più alta a (15a), dimostrando sensibilità alla struttura sintattica nonostante siano addestrate solo su obiettivo di next-word prediction senza supervisione sintattica esplicita.

Recursive Neural Networks per Composizione Sintattica

Per modellare esplicitamente la struttura sintattica gerarchica, sono stati sviluppati modelli neurali ricorsivi che operano su alberi piuttosto che sequenze. Le Recursive Neural Networks (RecNN) computano rappresentazioni vettoriali di costituenti sintattici in modo bottom-up, combinando ricorsivamente le rappresentazioni dei nodi figli mediante una funzione di composizione.

Nel modello più semplice, due vettori rappresentazionali (per i figli sinistro e destro) sono concatenati e trasformati mediante una rete feed-forward:

$h_{parent} = \tanh(W [h_{left}, h_{right}] + b)$

dove:

- $h_{left}, h_{right} \in \mathbb{R}^d$ sono le rappresentazioni dei costituenti figli
- $W \in \mathbb{R}^{d \times 2d}$ è la matrice di composizione
- $b \in \mathbb{R}^d$ è il bias

- h_{parent} ∈ ℝ^d è la rappresentazione del costituente genitore

Questa funzione di composizione è applicata ricorsivamente partendo dalle foglie (parole, rappresentate da word embeddings) risalendo fino alla radice dell'albero (rappresentazione dell'intera frase).

Socher et al. (2011, 2013) hanno esteso questo approccio introducendo le **Recursive Neural Tensor Networks (RNTN)**, che utilizzano una forma bilineare tensoriale per catturare interazioni più ricche tra i costituenti:

h_{parent} = tanh(h_{left}^T T h_{right} + W [h_{left}, h_{right}] + b)

dove T ∈ ℝ^{d×d×d} è un tensore tridimensionale (d slice 2D). Il termine h_{left}^T T h_{right} calcola:

h_{left}^T T h_{right} = [h_{left}^T T^{[1]} h_{right}, h_{left}^T T^{[2]} h_{right}, ..., h_{left}^T T^{[d]} h_{right}]

catturando interazioni moltiplicative tra ogni dimensione di h_{left} e h_{right}.

Le RNTN possono essere addestrate per vari task:

1. **Sentiment classification**: Dato albero sintattico, predire polarità (positivo/negativo/neutrale) a ogni nodo. Il modello impara rappresentazioni che catturano composizione semantica compositiva (es. negazione "not good" inverte polarità di "good").

2. **Parsing**: Assegnare punteggi a possibili alberi sintattici per una frase, scegliendo l'albero con punteggio massimo. Il punteggio può essere basato sulla norma della rappresentazione finale o su una funzione classificatrice.

3. **Paraphrase detection**: Date due frasi con alberi sintattici, computare similarità tra le rappresentazioni radice per determinare se sono parafrasi.

Un vantaggio delle RecNN è che le rappresentazioni intermedie catturano significato composizionale dei costituenti, permettendo reasoning a livello di sintagma. Tuttavia, richiedono alberi sintattici come input (devono essere forniti o parsati preventivamente), e sono computazionalmente più costose delle RNN sequenziali.

Modelli Neurali per il Parsing

I modelli neurali possono essere utilizzati non solo per modellare language ma anche per il parsing stesso, ovvero predire strutture sintattiche. L'approccio **transition-based parsing** formula il parsing come sequenza di decisioni (azioni) che costruiscono incrementalmente l'albero sintattico. Il parser mantiene una configurazione che include:

- **Stack**: Contenente parole/costituenti parzialmente processati
- **Buffer**: Contenente parole non ancora processate

- **Arcs**: Insieme di archi di dipendenza costruiti fino a ora (per dependency parsing)

A ogni passo, il parser sceglie un'azione tra un set finito (tipicamente 3 azioni base):

- **SHIFT**: Sposta la prima parola dal buffer allo stack
- **REDUCE-LEFT**: Crea arco di dipendenza da elemento in cima allo stack a elemento sottostante, rimuove elemento sottostante
- **REDUCE-RIGHT**: Come REDUCE-LEFT ma direzione inversa

Il parser inizia con configurazione (stack=[ROOT], buffer=frase, arcs={}), applica sequenza di azioni, e termina quando buffer è vuoto. La sequenza di azioni determina univocamente l'albero risultante.

Il modello neurale funziona come policy: dato lo stato corrente (configurazione), predice quale azione eseguire. Chen e Manning (2014) hanno introdotto un parser neurale transition-based che rappresenta la configurazione mediante concatenazione di embeddings:

- Top 3 elementi dello stack: s_1, s_2, s_3
- Top 3 elementi del buffer: b_1, b_2, b_3
- Dipendenti sinistri/destri degli elementi stack: lc_1, rc_1, lc_2, rc_2

La rappresentazione è:

x = [e(s_1), e(s_2), e(s_3), e(b_1), e(b_2), e(b_3), e(lc_1), e(rc_1), ...]

dove e(w) ∈ ℝ^d è l'embedding della parola w (appreso). Questa rappresentazione passa attraverso una rete feed-forward:

h = ReLU(W_1 x + b_1) scores = W_2 h + b_2

dove scores ∈ ℝ^{|A|} contiene un punteggio per ciascuna azione possibile, e |A| tipicamente è 3 (SHIFT, REDUCE-LEFT, REDUCE-RIGHT). La distribuzione di probabilità è:

P(action | configuration) = softmax(scores)

Il parser sceglie l'azione con probabilità massima (greedy) o usa beam search per mantenere k-best candidati.

L'addestramento avviene mediante supervised learning su treebank: per ogni frase con albero gold, si computa la sequenza gold di azioni che costruisce quell'albero, poi si massimizza:

L = -∑_t log P(action_t^{gold} | configuration_t)

sommando su tutti i time steps t di tutte le frasi di training.

Il parser di Chen e Manning ha ottenuto accuratezza comparabile a parser tradizionali basati su features manuali, ma con velocità drasticamente superiore (1000+ frasi/secondo) grazie alla semplicità del modello (solo due layer fully-connected). Successivi miglioramenti hanno

utilizzato LSTM bidirezionali per codificare le parole nello stack e buffer, raggiungendo accuratezza state-of-the-art (UAS ≈ 94% su Penn Treebank, dove UAS = Unlabeled Attachment Score misura frazione di archi di dipendenza predetti correttamente ignorando le label).

Transformer e Modelli Linguistici Pre-addestrati

L'architettura **Transformer**, introdotta da Vaswani et al. (2017) nel paper seminale "Attention is All You Need", ha ulteriormente rivoluzionato la modellazione del linguaggio sostituendo la ricorrenza con meccanismi di self-attention che permettono di catturare dipendenze arbitrariamente distanti in modo efficiente e completamente parallelizzabile. Questo ha reso possibile l'addestramento di modelli su dataset enormi (trilioni di token) e con miliardi di parametri, portando a capacità linguistiche senza precedenti.

Meccanismo di Self-Attention

Il cuore del Transformer è il meccanismo di **scaled dot-product attention**. Data una sequenza di vettori di input X = [x_1, x_2, ..., x_n] ∈ ℝ^{n×d}, l'attention computa una rappresentazione contestualizzata di ciascun elemento come media pesata di tutti gli elementi, dove i pesi dipendono dalla similarità tra gli elementi.

Formalmente, l'input viene proiettato linearmente in tre spazi: query (Q), key (K), e value (V):

Q = X W_Q (dove W_Q ∈ ℝ^{d×d_k}) K = X W_K (dove W_K ∈ ℝ^{d×d_k}) V = X W_V (dove W_V ∈ ℝ^{d×d_v})

L'attention è quindi computata come:

Attention(Q, K, V) = softmax(Q K^T / √d_k) V

Analizzando passo per passo:

1. **Q K^T**: Prodotto matriciale ∈ ℝ^{n×n}. L'elemento (i,j) è q_i^T k_j, la similarità (dot product) tra query i e key j.

2. **Scaling**: Divisione per √d_k riduce la magnitudine, prevenendo che il softmax saturi (argomenti troppo grandi → gradiente nullo).

3. **Softmax per righe**: Per ogni posizione i, softmax(Q K^T / √d_k)_{i,:} è una distribuzione di probabilità su tutte le posizioni j, indicando quanto la posizione i "attende a" ogni posizione j.

4. **Prodotto con V**: Il risultato ∈ ℝ^{n×d_v} è una media pesata dei value vectors, dove i pesi sono le attention probabilities.

Il risultato finale per la posizione i è:

output_i = ∑*{j=1}^{n} α*{ij} v_j

dove α_{ij} = softmax(q_i^T k_j / √d_k) è il peso di attention della posizione i verso j.

Intuitivamente, il meccanismo permette a ogni posizione di "guardare" tutte le altre posizioni e aggregare informazione da quelle rilevanti. La rilevanza è determinata dinamicamente mediante la similarità query-key.

Multi-Head Attention

In pratica, i Transformer utilizzano **multi-head attention**: il meccanismo di attention viene applicato h volte in parallelo con diverse matrici di proiezione, e i risultati vengono concatenati:

MultiHead(X) = Concat(head_1, head_2, ..., head_h) W_O

dove:

head_i = Attention(X W_Q^{(i)}, X W_K^{(i)}, X W_V^{(i)})

e W_O ∈ ℝ^{(h d_v)×d} proietta il risultato concatenato nello spazio originale.

Tipicamente si usa h = 8 o 12, con d_k = d_v = d/h, in modo che la dimensionalità totale rimanga d. Ogni head può potenzialmente imparare a catturare pattern di dipendenza diversi: alcuni head potrebbero specializzarsi in dipendenze sintattiche (soggetto-verbo), altri in relazioni semantiche (anafora), altri ancora in fenomeni di lungo raggio.

Architettura Completa del Transformer

Un blocco Transformer completo include:

1. **Multi-head self-attention**
2. **Residual connection + layer normalization**
3. **Feed-forward network (position-wise)**
4. **Residual connection + layer normalization**

Matematicamente:

X' = LayerNorm(X + MultiHead(X))
X'' = LayerNorm(X' + FFN(X'))

dove FFN è una rete feed-forward applicata indipendentemente a ogni posizione:

FFN(x) = max(0, x W_1 + b_1) W_2 + b_2

tipicamente con W_1 ∈ ℝ^{d×4d} (espansione), W_2 ∈ ℝ^{4d×d} (contrazione), e attivazione ReLU.

Il Transformer completo consiste di una stack di L blocchi (tipicamente L = 6-24 per modelli base, fino a 96+ per modelli molto grandi):

X^{(0)} = Embed(input) + PositionalEncoding
X^{(ℓ)} = TransformerBlock^{(ℓ)}(X^{(ℓ-1)}) per ℓ = 1, ..., L
output = X^{(L)}

Il PositionalEncoding è essenziale perché l'attention è permutation-invariant (non codifica ordine intrinsecamente). Si utilizzano encoding sinusoidali:

PE(pos, 2i) = sin(pos / 10000^{2i/d}) PE(pos, 2i+1) = cos(pos / 10000^{2i/d})

che permettono al modello di apprendere a utilizzare informazione posizionale relativa.

BERT: Bidirectional Encoder Representations from Transformers

BERT (Devlin et al., 2019) è un modello basato su Transformer encoder pre-addestrato con un obiettivo di **masked language modeling (MLM)**. L'idea è mascherare casualmente una percentuale (tipicamente 15%) dei token in una sequenza e addestrare il modello a predire i token mascherati basandosi sul contesto bidirezionale:

Input: "The cat [MASK] on the [MASK]" Output: P("sat"|context), P("mat"|context)

L'obiettivo di training è:

L_{MLM} = -$\mathbb{E}$_{sequenze} [$\sum_{i \in masked} \log P(w_i \mid w_{\setminus i})$]

dove w_{\i} denota la sequenza con w_i mascherato.

Questo obiettivo forza il modello a costruire rappresentazioni contestuali profonde di ogni token, integrando informazione da sinistra e destra. A differenza di language models autoregressivi (come GPT) che vedono solo contesto sinistro, BERT ha accesso bidirezionale, potenzialmente permettendo rappresentazioni più ricche.

BERT viene pre-addestrato su corpus enormi (Wikipedia + BookCorpus ≈ 3,3 miliardi di parole per BERT-base), poi fine-tuned su task specifici (classificazione, NER, QA, etc.) con pochi esempi etichettati. Il pre-training cattura conoscenza linguistica generale (inclusa sintassi), che viene riutilizzata per vari task downstream.

GPT: Generative Pre-Training

GPT (Generative Pre-trained Transformer, Radford et al. 2018-2020) è una famiglia di modelli basati su Transformer decoder pre-addestrati con standard autoregressive language modeling:

L_{AR} = -$\mathbb{E}$_{sequenze} [$\sum_{i=1}^{n} \log P(w_i \mid w_{<i})$]

dove w_{<i} = (w_1, ..., w_{i-1}) è il contesto precedente.

A differenza di BERT che usa encoder bidirezionale, GPT usa decoder con **masked self-attention**: al tempo i, l'attention può guardare solo posizioni ≤ i (contesto precedente), non future. Questo è implementato mediante una mask triangolare superiore nella matrice di attention scores.

GPT-3 (Brown et al., 2020) scala a dimensioni senza precedenti:

- 175 miliardi di parametri
- Training su ≈ 500 miliardi di token (≈ 45TB di testo)
- 96 layer Transformer, 96 attention heads, dimensionalità d = 12288

GPT-3 ha dimostrato capacità sorprendenti di **few-shot learning**: fornendo pochi esempi nel prompt (in-context learning), il modello può adattarsi a task nuovi senza fine-tuning. Per esempio:

Translate English to French:
"Hello" → "Bonjour"
"Goodbye" → "Au revoir"
"How are you?" → [GPT-3 genera "Comment allez-vous?"]

Questa capacità suggerisce che modelli sufficientemente grandi con training sufficiente internalizzano pattern linguistici estremamente generali, applicabili flessibilmente a contesti nuovi.

Cattura di Conoscenza Sintattica in Modelli Pre-addestrati

Studi recenti hanno investigato sistematicamente in che misura BERT, GPT e modelli simili catturano conoscenza sintattica. Il **BLiMP benchmark** (Benchmark of Linguistic Minimal Pairs, Warstadt et al. 2020) consiste di 67.000 coppie minimali di frasi che testano vari fenomeni sintattici. Ogni coppia include una frase accettabile e una minimamente diversa ma inaccettabile:

Agreement: (16a) The author who the senators admire writes political satire. ✓ (16b) *The author who the senators admire write political satire. ✗

Island effects: (17a) What do you think that John bought? ✓ (17b) *What do you wonder whether John bought? ✗

Binding: (18a) John thinks that Mary likes himself. ✗ (18b) John thinks that Mary likes herself. ✓

NPI licensing: (19a) No student has ever been to Paris. ✓ (19b) *Some student has ever been to Paris. ✗

Per ciascuna coppia, il modello linguistico deve assegnare probabilità maggiore alla frase accettabile. L'accuratezza è:

accuracy = (1/N) × ∑_{(s_1,s_2) ∈ pairs} 𝟙[P(s_1) > P(s_2)]

dove $\mathbb{1}$ è la funzione indicatore.

I risultati per GPT-3 e altri modelli grandi mostrano performance variabile ma generalmente elevata. L'accuratezza complessiva si attesta tra l'80% e il 90% su molti fenomeni testati, con performance particolarmente alta, superiore al 90%, su fenomeni locali come l'agreement semplice e l'anafora pronominale in domini ristretti. Questi risultati indicano che i modelli linguistici pre-addestrati hanno effettivamente internalizzato pattern sintattici sistematici, non solo sequenze superficiali di parole. Tuttavia, la performance degrada significativamente per fenomeni che richiedono dipendenze gerarchiche complesse o conoscenza di vincoli astratti.

Per esempio, sui fenomeni di island constraints, la performance di GPT-3 mostra pattern interessanti di gradazione che rispecchiano le intuizioni dei parlanti nativi. Le whether-islands, che producono giudizi di inaccettabilità molto forti nei parlanti umani, vengono identificate correttamente dal modello nell'85-88% dei casi. Al contrario, le adjunct islands, che nei giudizi umani producono effetti più sfumati e context-dependent, vengono identificate solo nel 65-72% dei casi. Questa correlazione tra severità percepita della violazione e performance del modello suggerisce che le rappresentazioni apprese catturano genuinamente aspetti della struttura sintattica graduata piuttosto che semplici pattern lessicali superficiali.

Analogamente, sui fenomeni di Negative Polarity Items, GPT-3 distingue contesti di licensing appropriati da contesti inappropriati con accuratezza del 78%, commettendo errori sistematici proprio in quei casi borderline che anche i parlanti nativi trovano problematici. Espressioni come "ever" in inglese richiedono un licensor downward-entailing nel loro dominio sintattico. La frase "No student has ever been to Paris" è accettabile perché "no" è downward-entailing, mentre la frase con "some student has ever been to Paris" è inaccettabile. Il modello cattura questa distinzione nella maggioranza dei casi, ma mostra incertezza in costruzioni dove l'ambito del licensor è ambiguo o dove intervengono fattori pragmatici.

Gli studi di probing, metodologia sperimentale sviluppata per investigare quali informazioni linguistiche sono codificate nelle rappresentazioni interne dei modelli neurali, hanno fornito evidenze convergenti. Il probing consiste nell'addestrare classificatori semplici (tipicamente regressione logistica lineare) a predire proprietà linguistiche specifiche utilizzando come input le rappresentazioni hidden state del modello. L'idea è che se una proprietà linguistica P può essere predetta accuratamente da uno stato nascosto h mediante un classificatore lineare, allora l'informazione su P deve essere codificata esplicitamente in h in forma linearmente separabile.

Hewitt e Manning nel 2019 hanno introdotto il concetto di structural probes, classificatori progettati per testare se le rappresentazioni di BERT codificano alberi sintattici di dipendenza. Hanno definito una metrica di distanza nello spazio delle rappresentazioni che dovrebbe correlare con la distanza sintattica nell'albero di dipendenza. Formalmente, data una matrice di trasformazione B, la distanza sintattica predetta tra le parole i e j nella frase è definita come la distanza euclidea tra le loro rappresentazioni trasformate:

$$d_{syntax}(i,j) = ||B h_i - B h_j||_2$$

dove h_i e h_j sono le rappresentazioni contestuali di BERT per le posizioni i e j. La matrice B viene appresa minimizzando la differenza tra d_syntax(i,j) e la distanza effettiva nell'albero di dipendenza gold. Il risultato sorprendente è che esiste effettivamente una trasformazione lineare B tale che le distanze euclidee risultanti correlano fortemente con le distanze sintattiche, con coefficiente di correlazione di Spearman ρ superiore a 0,80 su vari linguaggi testati.

Inoltre, Hewitt e Manning hanno testato se le rappresentazioni codificano la profondità sintattica dei token nell'albero. Definendo la profondità predetta come la norma della rappresentazione trasformata:

depth(i) = ||B h_i||_2

hanno trovato correlazione significativa (ρ maggiore di 0,75) con la profondità effettiva nell'albero gold. Questi risultati dimostrano che BERT, addestrato solo con masked language modeling senza alcuna supervisione sintattica esplicita, ha sviluppato rappresentazioni che codificano struttura sintattica gerarchica in modo sistematico e geometricamente consistente.

Studi successivi hanno esplorato layer-wise specialization nelle rappresentazioni di BERT e GPT. Analizzando come la performance di probing varia attraverso i layer del modello, si osserva un pattern sistematico: i layer inferiori codificano principalmente informazione lessicale e morfologica, i layer intermedi codificano sintassi locale e dipendenze a media distanza, mentre i layer superiori codificano relazioni semantiche e pragmatiche più astratte. Quantitativamente, sui task di part-of-speech tagging, i migliori probe utilizzano rappresentazioni dal layer 3-5 di un modello a 12 layer, raggiungendo accuratezza del 96-98%. Sui task di dependency parsing, i migliori probe utilizzano layer 6-8, raggiungendo Unlabeled Attachment Score di 85-90%. Sui task semantici come semantic role labeling, i migliori probe utilizzano layer 9-11.

Questa stratificazione funzionale emerge spontaneamente dall'addestramento e riflette una gerarchia di astrazione linguistica: dal lessico alla sintassi alla semantica. Il fatto che questa organizzazione non sia imposta architetturalmente ma emerga dall'ottimizzazione dell'obiettivo di language modeling suggerisce che rappresenta una soluzione naturale ed efficiente al problema della predizione linguistica.

Tuttavia, l'interpretazione dei risultati di probing richiede cautela metodologica. Il fatto che un classificatore lineare possa estrarre informazione sintattica dalle rappresentazioni non implica necessariamente che il modello utilizzi questa informazione in modo causale per le sue predizioni. Un dibattito metodologico importante riguarda la distinzione tra informazione correlazionale e informazione funzionale. Le rappresentazioni potrebbero codificare struttura sintattica come sottoprodotto di regolarità statistiche apprese, senza che questa struttura giochi ruolo causale nel behavior del modello.

Per affrontare questa questione, sono stati sviluppati intervention methods che modificano direttamente le rappresentazioni interne del modello in modo chirurgico e osservano l'effetto sulle predizioni. Per esempio, Giulianelli et al. nel 2018 hanno manipolato le rappresentazioni di subject number in LSTM language models, invertendo artificialmente il numero grammaticale nella rappresentazione del soggetto in frasi con agreement a lunga

distanza. Se il modello utilizza causalmente questa informazione per la predizione del verbo, l'intervento dovrebbe causare errori di agreement sistematici. I risultati confermano questa predizione: alterare la rappresentazione del numero del soggetto causa aumento significativo di errori di agreement, dimostrando utilizzo causale dell'informazione sintattica.

Analogamente, studi di counterfactual interventions su BERT hanno mostrato che alterare le rappresentazioni di dipendenze sintattiche specifiche influenza predizioni downstream che dipendono da quelle strutture. Questo fornisce evidenza più forte che i modelli non solo codificano sintassi passivamente ma la utilizzano attivamente nel processamento.

Nonostante queste capacità impressionanti, i modelli neurali mostrano anche limitazioni sistematiche che rivelano differenze qualitative rispetto alla competenza linguistica umana. Un fenomeno particolarmente interessante è la sensitivity ridotta a violazioni sintattiche che richiedono ragionamento composizionale complesso. Marcus e Davis nel 2019 hanno documentato che GPT-2, nonostante performance generalmente alta, fallisce sistematicamente su costruzioni che richiedono applicazione ricorsiva di regole sintattiche a profondità arbitraria.

Consideriamo la capacità di processare center-embedding annidati. Come discusso precedentemente, frasi come "The rat the cat the dog chased caught ate the cheese" sono grammaticalmente ben formate ma praticamente incomprensibili per gli umani a causa di limitazioni di memoria di lavoro. I modelli neurali mostrano pattern diverso: GPT-3 può processare embeddings tripli con performance relativamente buona, assegnando loro probabilità non trascurabile e continuazioni semanticamente appropriate. Questo suggerisce che i vincoli di processamento che limitano gli umani non si applicano allo stesso modo ai modelli neurali, che possono mantenere stack arbitrariamente profondi di dipendenze incomplete.

Tuttavia, quando la profondità di embedding aumenta ulteriormente, anche i modelli neurali mostrano degradazione progressiva. La performance su quadruplo embedding è notevolmente peggiore, e con quintuplo embedding i modelli producono output essenzialmente casuale. La differenza importante è che questa degradazione è graduale nei modelli neurali, mentre negli umani è più categoriale: già a triplo embedding la comprensione collassa drammaticamente.

Un'altra limitazione sistematica riguarda la generalizzazione composizionale. I modelli addestrati su distribuzioni di training che includono costruzioni fino a una certa complessità strutturale spesso falliscono quando testati su costruzioni più complesse ma composizionalmente derivabili dalle strutture viste. Per esempio, un modello addestrato su frasi con al massimo 2 relative clause annidiate può avere difficoltà con 3 relative annidiate, anche se la regola ricorsiva è la stessa. Questo contrasta con la competenza umana, dove l'applicabilità ricorsiva illimitata delle regole sintattiche è principio fondamentale, anche se la performance effettiva è limitata da fattori di memoria e processamento.

Lake e Baroni nel 2018 hanno proposto il SCAN benchmark per testare generalizzazione composizionale in reti neurali. SCAN richiede traduzione da comandi linguistici a sequenze di azioni, dove i comandi hanno struttura composizionale sistematica. Per esempio, "jump twice" si traduce in "JUMP JUMP", "run and jump" in "RUN JUMP", e queste operazioni si possono combinare ricorsivamente. I modelli neurali sequence-to-sequence standard

falliscono drammaticamente quando testati su combinazioni non viste di operazioni primitive viste separatamente, raggiungendo accuratezza sotto il 20% in alcuni split del dataset.

Questo suggerisce che mentre i modelli neurali possono approssimare competenza sintattica entro la distribuzione di training, mostrano fragilità nella generalizzazione sistematica richiesta dalla vera competenza compositiva. La loro conoscenza sintattica è più simile a un vasto database di pattern memorizzati con capacità di interpolazione, piuttosto che a un sistema di regole combinatorie ricorsive.

Gradienza Sintattica e Architettura del Sistema di Processamento

Le evidenze empiriche sulla gradienza sintattica sollevano questioni profonde riguardo all'architettura cognitiva del sistema di processamento linguistico. Se la grammaticalità non è dicotomica ma graduale, come dovremmo concepire il rapporto tra competenza e performance, tra sistema rappresentazionale e sistema computazionale?

Una prospettiva influente è quella dei constraint-based models of sentence processing, sviluppati principalmente da MacDonald, Pearlmutter, Seidenberg e altri negli anni 1990. Secondo questo approccio, il processamento sintattico non coinvolge l'applicazione seriale di regole categoriali, ma piuttosto la soddisfazione parallela di vincoli multipli e graduali che derivano da varie fonti di informazione: probabilità lessicali, preferenze strutturali, plausibilità semantica, prosodia, contesto discorsivo.

Ogni vincolo contribuisce un peso probabilistico a favore o contro un'analisi sintattica particolare, e questi pesi si compongono (tipicamente in modo additivo in scala log) per determinare la preferenza complessiva. Formalmente, la probabilità di un'analisi A data una stringa w può essere scomposta come:

$$P(A|w) \propto \exp(\sum_i \lambda_i C_i(A,w))$$

dove $C_i(A,w)$ è il valore del vincolo i per l'analisi A della stringa w, e λ_i è il peso del vincolo (parametro appreso dall'esperienza). Questa è esattamente la formulazione di un modello log-lineare, e infatti i constraint-based models possono essere visti come precursori teorici dei modelli discriminativi discussi precedentemente.

I vincoli possono essere soft (graduali) piuttosto che hard (categoriali). Per esempio, il vincolo "preferisci attachment basso di PP" non esclude categorialmente l'attachment alto, ma assegna penalità graduata. Similmente, vincoli semantici come "preferisci interpretazioni plausibili" operano su scala continua di plausibilità piuttosto che dicotomia plausibile-implausibile.

Evidenze psicolinguistiche da studi di reading time e eye-tracking supportano questa concezione. Quando i partecipanti leggono frasi ambigue, i tempi di lettura mostrano effetti graduali di difficoltà di processamento che correlano con la forza dei vincoli violati piuttosto che mostrare pattern categoriali. Per esempio, nella classica ambiguità di attachment:

(20) The spy saw the cop with the binoculars.

dove "with the binoculars" può attaccarsi al VP (lo spy usava binocoli per vedere) o all'NP "the cop" (il cop possedeva binocoli), i tempi di lettura nella regione disambiguante riflettono il grado di sorpresa rispetto all'interpretazione preferita. Se il contesto precedente rende molto probabile un'interpretazione, la forzatura all'interpretazione alternativa causa rallentamento proporzionale alla divergenza di probabilità.

Questo è catturato dall'effetto di surprisal, proposto da Hale nel 2001 e Levy nel 2008 come linking hypothesis tra probabilità linguistiche e difficoltà di processamento. Il surprisal di una parola w_i dato il contesto precedente w_{1:i-1} è definito come il logaritmo negativo della probabilità condizionale:

surprisal(w_i) = -log P(w_i | w_{1:i-1})

La teoria predice che il tempo di lettura su w_i dovrebbe essere proporzionale al suo surprisal:

RT(w_i) = α + β × surprisal(w_i) + ε

dove α è l'intercetta, β è il coefficiente di surprisal, e ε è rumore residuo. Questa predizione è stata testata estensivamente utilizzando language models per stimare P(w_i | w_{1:i-1}), e i risultati mostrano correlazioni significative tra surprisal e reading time, con β tipicamente nell'ordine di 5-15 millisecondi per unità di surprisal (nat o bit, a seconda della base del logaritmo).

Inoltre, studi hanno dimostrato che non solo il surprisal lessicale (probabilità della parola specifica) ma anche il surprisal sintattico (probabilità della struttura sintattica) contribuisce indipendentemente ai tempi di processamento. Roark et al. nel 2009 hanno mostrato che in modelli che includono sia lexical surprisal sia syntactic surprisal (calcolato come -log P(analisi sintattica | contesto)), entrambe le componenti contribuiscono varianza significativa nella predizione dei reading time. Questo supporta l'idea che il processamento sintattico è intrinsecamente probabilistico: il parser umano è sensibile alle probabilità relative di strutture sintattiche alternative.

Un fenomeno particolarmente interessante è l'effetto di syntactic satiation o exposure effect sulle violazioni di isola. Snyder nel 2000 ha documentato che l'esposizione ripetuta a violazioni di island constraints durante un esperimento causa graduale aumento di acceptability nei giudizi dei partecipanti. Per esempio, se i partecipanti giudicano ripetutamente estrazioni da whether-islands, inizialmente giudicate con severità media di 2,5 su scala Likert a 7 punti, dopo 20-30 esposizioni i giudizi medi aumentano a 3,5-4,0.

Questo effetto è specifico al tipo di isola: l'esposizione a whether-islands non migliora significativamente i giudizi su complex NP islands. Inoltre, l'effetto mostra gradiente di severità: le violazioni inizialmente giudicate meno severe mostrano satiation più rapida ed estesa rispetto a violazioni inizialmente più severe. Le whether-islands, tipicamente le più severe, mostrano satiation modesta, mentre adjunct islands mostrano satiation sostanziale.

L'interpretazione teorica di questi fenomeni è dibattuta. Una possibilità è che riflettano adattamento delle aspettative probabilistiche: dopo esposizione ripetuta a costruzioni rare, il parser aggiorna le sue stime di probabilità mediante meccanismi di apprendimento

bayesiano. Se P(struttura|contesto) è la probabilità prior di una struttura, l'osservazione ripetuta di quella struttura dovrebbe aumentare la sua probabilità posterior secondo la regola di Bayes:

$$P(\text{struttura}|\text{osservazioni}) \propto P(\text{osservazioni}|\text{struttura}) \times P(\text{struttura})$$

Dopo n osservazioni della struttura, la posterior tende a essere superiore alla prior, riflettendo evidenza accumulata che la struttura è più comune di quanto inizialmente stimato.

Un'interpretazione alternativa è che la satiation rifletta habituazione percettuale: la risposta del sistema di processamento a stimoli ripetuti si attenua gradualmente, non perché le rappresentazioni sottostanti cambino, ma perché i meccanismi di rilevazione della novità si desensibilizzano. Questa interpretazione predice che l'effetto dovrebbe essere temporaneo e dissiparsi dopo delay, mentre l'interpretazione basata su apprendimento predice effetti più duraturi. Gli studi hanno trovato evidenze miste: alcuni mostrano persistenza parziale dell'effetto dopo intervalli, altri mostrano dissipazione rapida.

Construction Grammar e Modelli Usage-Based della Competenza Linguistica

Le evidenze sulla gradienza sintattica hanno motivato lo sviluppo di framework teorici alternativi al generativismo classico, che concepiscono la competenza linguistica in termini fondamentalmente diversi. La Construction Grammar, nelle sue varie incarnazioni (Cognitive Grammar di Langacker, Construction Grammar di Fillmore e Kay, Radical Construction Grammar di Croft), propone che la conoscenza linguistica consista in un inventario di costruzioni, ovvero accoppiamenti convenzionalizzati di forma e significato che variano in dimensione, specificità e produttività.

Una costruzione può essere altamente specifica e idiomatica, come "kick the bucket" con significato "morire", oppure astratta e schematica come la costruzione ditransitiva [Subj V Obj1 Obj2], che esprime transfer causato. Le costruzioni formano una rete tassonomica con relazioni di ereditarietà: costruzioni specifiche ereditano proprietà da costruzioni più generali, ma possono anche mostrare idiosincrasie non derivabili dalla costruzione madre.

Questo approccio dissolve la distinzione netta tra lessico e sintassi postulata dal generativismo: entrambi sono visti come punti lungo un continuum di costruzioni che varia in schematicità. Morfemi e parole sono costruzioni massimamente specifiche, mentre schemi sintattici astratti sono costruzioni massimamente generali, ma non vi è discontinuità qualitativa.

La gradienza sintattica emerge naturalmente in questo framework attraverso il concetto di entrenchment e productivity. Le costruzioni non sono classificate binariamente come "nella lingua" vs "fuori dalla lingua", ma piuttosto lungo dimensioni continue di quanto sono entrenched nella competenza del parlante (funzione della frequenza di esposizione) e quanto sono produttive (applicabili a range ampio di contesti).

Una costruzione altamente entrenched e produttiva genera istanze con alta acceptability. Una costruzione poco entrenched (rara) o poco produttiva (limitata a contesti specifici)

genera istanze con acceptability ridotta quando usata in contesti atipici. Una sequenza che non corrisponde a nessuna costruzione esistente ma può essere analizzata per analogia con costruzioni simili riceve acceptability intermedia, proporzionale alla similarità.

Formalmente, alcuni modelli usage-based hanno proposto che la probabilità di una costruzione C applicata con lessemi specifici L in contesto K può essere modellata come:

P(C, L, K) = freq(C) × typfreq(C) × similarity(L, prototypical_fillers(C)) × fit(K, typical_contexts(C))

dove freq(C) è la frequenza token di C (quante volte è stata incontrata), typfreq(C) è la frequenza type (con quanti lessemi diversi), similarity misura la similarità semantica tra i lessemi L e i filler prototipici della costruzione, e fit misura l'appropriatezza del contesto K.

Questa formulazione cattura vari effetti empirici documentati. Per esempio, la costruzione causativa in inglese "make X do Y" è più accettabile con verbi emotivi o percettivi come "make someone laugh/cry/see" rispetto a verbi di azione fisica come "make someone walk/run", riflettendo le distribuzioni di frequenza: i verbi emotivi/percettivi sono fillers più tipici basati su evidenza corpus. Il modello predice correttamente che "make someone contemplate" (verbo cognitivo raro in questa costruzione) riceve acceptability intermedia: il verbo è semanticamente simile ai fillers tipici ma ha frequenza bassa.

Studi corpus-based hanno quantificato queste predizioni. Goldberg et al. nel 2004 hanno analizzato la costruzione ditransitiva in inglese, codificando per ogni verbo la frequenza di occorrenza nella costruzione e misurando la similarità semantica tra verbi mediante analisi distribuzionale automatica. Hanno trovato che la facilità con cui i parlanti accettano verbi nuovi nella costruzione ditransitiva (testata mediante rating experiment) correlava significativamente con entrambe le misure: r = 0,68 per frequenza, r = 0,53 per similarità semantica con verbi altamente frequenti nella costruzione.

Inoltre, la produttività di una costruzione può essere quantificata mediante misure derivate dalla distribuzione type-token. Una misura influente è il Productivity Index proposto da Baayen nel 1992 nel contesto morfologico ma applicabile anche alla sintassi:

P = n_1 / N

dove n_1 è il numero di types che occorrono esattamente una volta (hapax legomena) e N è il numero totale di tokens. L'idea è che costruzioni produttive generano continuamente istanze con lessemi nuovi, risultando in alta proporzione di hapax. Costruzioni non produttive sono limitate a set ristretto di collocazioni, risultando in bassa proporzione di hapax (gli stessi types ricorrono ripetutamente).

Per esempio, la costruzione way-construction in inglese "X verb one's way PP" come in "He elbowed his way through the crowd" ha productivity index relativamente basso (P minore di 0,15 in analisi corpus), indicando che è semi-produttiva: accetta verbi nuovi ma con restrizioni. Al contrario, costruzioni core come la transitiva semplice hanno P maggiore di 0,40, indicando alta produttività.

Questi approcci usage-based e basati su costruzioni offrono un account naturale della gradualità: l'acceptability è funzione continua di fattori multipli (frequenza, tipicalità, similarità, produttività) che si compongono probabilisticamente. Non vi è necessità di postulare dicotomia categoriale tra grammaticale e agrammaticale; piuttosto, tutte le strutture ricevono gradi di acceptability determinati dalla loro conformità con pattern estratti dall'esperienza linguistica.

Tuttavia, questi approcci affrontano sfide teoriche ed empiriche. Una sfida è fornire account esplicito della creatività linguistica: come possono i parlanti produrre e comprendere frasi mai incontrate prima, combinando costruzioni in modi nuovi? Il generativismo risponde postulando regole ricorsive che generano infiniti output da mezzi finiti. I modelli usage-based rispondono invocando analogia e generalizzazione schematica, ma i meccanismi computazionali precisi rimangono spesso sotto-specificati.

Un'altra sfida riguarda universali linguistici. Se la grammatica emerge dall'uso, come spiegare pattern cross-linguistici sistematici? Per esempio, soggetti tendono a precedere oggetti in maggioranza delle lingue del mondo; ordini SOV e SVO coprono oltre l'85% delle lingue. I generativisti spiegano questo mediante principi innati della Universal Grammar. I construction grammarians invocano fattori funzionali e cognitivi generali: iconicità (agente prima azione prima paziente), economia (minimizzare distanza tra elementi dipendenti), salienza percettiva (animati prima inanimati), che favorirebbero indipendentemente certi ordini.

Queste spiegazioni funzionali sono spesso plausibili ma difficili da testare rigorosamente. Inoltre, non è chiaro se spiegano l'intero range di universali documentati, alcuni dei quali sembrano arbitrari dal punto di vista funzionale. Per esempio, la Greenberg's Universal 20 afferma che in lingue VSO l'ordine Noun-Genitive è predominante. Non è ovvio quale pressione funzionale generale spiegherebbe questa correlazione specifica tra ordine verbale e ordine nominale.

Sintesi Teorica: Verso un Modello Integrato della Competenza Graduale

Le evidenze e le prospettive teoriche discusse convergono verso una concezione della competenza linguistica che differisce profondamente dal modello generativista classico. La competenza non è un sistema di regole categoriali che specifica l'insieme discreto delle strutture grammaticali, ma piuttosto un sistema probabilistico che assegna gradi di acceptability a strutture basandosi su molteplici fonti di informazione integrate.

Questo sistema incorpora conoscenza di vari tipi di regolarità linguistica organizzate gerarchicamente. A livello più basso, include conoscenza lessicale: quali morfemi esistono, le loro forme fonologiche, i loro significati, le loro proprietà combinatorie idiosincratiche. A livello intermedio, include conoscenza di pattern costruzionali di specificità variabile: collocazioni multi-word, frame argomentuali dei verbi, costruzioni sintattiche semi-produttive. A livello più alto, include schemi astratti molto generali che corrispondono approssimativamente alle "regole" della grammatica tradizionale: struttura di frase X-bar, dipendenze di agreement, vincoli su movimenti.

Ogni livello contribuisce vincoli che influenzano la formazione e l'interpretazione di enunciati. I vincoli non sono assoluti ma graduali, e possono interagire in modo complesso: violarsi reciprocamente, rinforzarsi, o applicarsi con forza variabile in contesti diversi. La competenza include anche conoscenza delle frequenze: quali strutture sono comuni o rare, quali combinazioni lessicali sono tipiche o atipiche.

La performance linguistica emerge dall'interazione in tempo reale di questo sistema di conoscenza con sistemi cognitivi generali: memoria di lavoro, attenzione, processi di decisione. Limitazioni di memoria spiegano difficoltà con embeddings profondi; preferenze di processamento spiegano fenomeni di garden path; fattori contestuali e pragmatici modulano interpretazione e acceptability.

Questa concezione dissolve la dicotomia netta tra competenza e performance postulata da Chomsky. La competenza stessa è probabilistica e graduale, riflettendo l'esperienza linguistica accumulata del parlante. La performance aggiunge ulteriori fattori di variabilità, ma non è qualitativamente diversa: entrambe coinvolgono vincoli multipli e graduali.

I modelli computazionali contemporanei, in particolare i grandi language models neurali, approssimano aspetti di questa competenza graduale. Addestrati su corpora enormi con obiettivo di predizione, internalizzano distribuzioni probabilistiche su strutture linguistiche che correlano significativamente con giudizi umani di acceptability. Tuttavia, mostrano anche differenze sistematiche: generalizzazione composizionale limitata, sensibilità ridotta a vincoli astratti, assenza di capacità meta-linguistiche.

Queste differenze rivelano che la competenza umana include componenti che vanno oltre pattern statistici estratti da esposizione. Probabilmente include bias induttivi innati che guidano l'apprendimento verso generalizzazioni appropriate, capacità di ragionamento esplicito su struttura linguistica, e integrazione profonda con sistemi concettuali e comunicativi. La sfida teorica è specificare precisamente questi componenti e come interagiscono con apprendimento statistico.

Un modello integrato promettente potrebbe combinare elementi di approcci diversi. Dal generativismo, mantiene l'idea che esistono principi strutturali generali che vincolano possibili grammatiche umane, ma concepisce questi principi come bias probabilistici piuttosto che vincoli categoriali. Dalla Construction Grammar, adotta l'idea che conoscenza linguistica include costruzioni di specificità variabile organizzate in rete, con proprietà emergenti dall'uso. Dai modelli neurali, incorpora meccanismi di apprendimento statistico che estraggono regolarità distribuzionali da input linguistico.

Formalmente, tale modello potrebbe essere specificato come distribuzione probabilistica gerarchica bayesiana. I livelli superiori della gerarchia codificano prior su tipi di strutture possibili (corrispondenti a Universal Grammar), i livelli intermedi codificano parametri specifici della lingua appresa (costruzioni e pattern), e i livelli inferiori codificano istanze lessicali specifiche. L'apprendimento procede mediante inferenza bayesiana: data esperienza linguistica, aggiorna la distribuzione posteriore sui parametri a tutti i livelli.

Questa formulazione cattura naturalmente fenomeni di gradualità: strutture consistenti con prior forti e supportate da evidenza frequente ricevono probabilità alta, strutture che violano prior o hanno scarsa evidenza ricevono probabilità bassa, e vi è continuum tra gli estremi.

Inoltre, permette learning a livelli multipli simultaneamente: affinamento di costruzioni specifiche mentre si mantengono principi generali, o modifica graduale di principi generali se evidenza è sufficientemente forte e sistematica.

Un vantaggio di questa prospettiva è che unifica fenomeni apparentemente disparati. La gradualità nei giudizi di acceptability, gli effetti di frequenza nel processamento, la variazione cross-linguistica entro range limitato, l'apprendimento linguistico dei bambini mediante esposizione, la creatività linguistica entro vincoli, possono tutti essere visti come manifestazioni di un sistema unico che combina apprendimento statistico con bias strutturali.

L'implementazione computazionale dettagliata di tale modello rimane sfida aperta. I modelli bayesiani gerarchici della sintassi proposti finora (per esempio, adaptor grammars, Bayesian PCFGs con prior gerarchi) catturano alcuni aspetti ma non la ricchezza completa della competenza umana. I grandi language models neurali catturano pattern complessi ma mancano di struttura esplicita e interpretabilità. La sintesi delle due tradizioni rappresenta direzione promettente per ricerca futura.

Implicazioni per la Metodologia Linguistica e la Raccolta di Dati

La dissoluzione della dicotomia grammaticale ha implicazioni profonde per la metodologia della ricerca linguistica. Se la grammaticalità è graduale, come dovrebbero i linguisti raccogliere e analizzare dati empirici? La pratica tradizionale di giudizi informali forniti dal linguista stesso o da piccoli gruppi di parlanti, codificati come simboli discreti (asterisco, punto interrogativo, check mark), è chiaramente inadeguata.

La linguistica sperimentale quantitativa offre strumenti più appropriati. Studi su scala ampia con campioni rappresentativi di parlanti, utilizzando scale graduate (Likert scales, magnitude estimation) e analisi statistiche rigorose, permettono di caratterizzare distribuzioni di acceptability e identificare fattori che influenzano giudizi. I metodi di regressione multipla consentono di disentanglare contributi di fattori diversi (lunghezza, frequenza, complessità strutturale, plausibilità semantica) alla varianza nei giudizi.

Tuttavia, anche i giudizi di acceptability raccolti sperimentalmente presentano limiti. Riflettono non solo competenza linguistica ma anche meta-conoscenza linguistica, attitudini prescrittive, effetti di task demand, e variabilità individuale sistematica. Due parlanti nativi della stessa varietà linguistica possono dare giudizi divergenti su costruzioni marginali, e tali divergenze possono riflettere differenze genuine nelle loro grammatiche (acquisite mediante esposizione a input diversi) piuttosto che rumore casuale.

Inoltre, alcuni fenomeni sintattici possono essere meglio investigati mediante metodi indiretti che misurano comportamento piuttosto che giudizi espliciti. Gli studi di sentence processing utilizzano tecniche come self-paced reading, eye-tracking durante lettura, e EEG/MEG per misurare correlati temporali del processamento. Questi metodi sono meno suscettibili a bias meta-linguistici e catturano aspetti della competenza che potrebbero non essere accessibili consciamente.

Per esempio, violazioni di agreement producono pattern caratteristici di ERP (Event-Related Potentials): la componente P600, un'onda positiva con picco attorno a 600 millisecondi dopo onset della violazione. L'ampiezza della P600 è modulata dalla severità della violazione: violazioni più gravi producono P600 di ampiezza maggiore. Questo fornisce misura neurofisiologica continua di degradazione grammaticale che complementa i giudizi comportamentali.

Analogamente, studi di priming sintattico misurano la tendenza a riutilizzare strutture sintattiche recentemente incontrate. Se un parlante produce o processa una costruzione passiva, è più probabile che produca un'altra passiva successivamente, anche con lessico completamente diverso. L'entità del priming riflette quanto la struttura è rappresentata astrattamente nella competenza. Strutture altamente produttive e entrenched mostrano priming robusto; costruzioni idiosincratiche lessicalizzate mostrano priming ridotto o assente.

L'integrazione di metodi multipli offre triangolazione: convergenza tra giudizi espliciti, misure di processamento online, analisi corpus, e modellazione computazionale aumenta confidenza nelle conclusioni teoriche. Divergenze tra metodi sono anche informative, rivelando che diversi aspetti del sistema linguistico possono essere accessibili mediante task diversi.

L'analisi corpus ha importanza particolare nell'era dei big data linguistici. Corpora di miliardi di parole (Google Books, Common Crawl, Wikipedia) permettono di stimare frequenze di costruzioni anche rare con precisione statistica adeguata. Inoltre, metodi di corpus linguistici possono identificare fenomeni di gradualità che sarebbero difficili da scoprire mediante introspezione.

Per esempio, Bresnan e collaboratori hanno utilizzato analisi corpus per identificare fattori multipli che influenzano l'alternanza dativa in inglese, scoprendo che animatezza, pronominalità, definitezza, e lunghezza degli argomenti contribuiscono tutte indipendentemente alla scelta tra forma double object e forma prepositional. Questi fattori formano gerarchia continua di preferenza piuttosto che regola categorica.

Tuttavia, i dati corpus presentano anche limiti. Riflettono performance, non direttamente competenza, e possono essere distorti da fattori stilistici, registri, e convenzioni di genere testuale. Costruzioni perfettamente grammaticali ma rare o formali possono avere frequenza molto bassa nei corpora informali. Inoltre, corpora non contengono informazione negativa: l'assenza di una costruzione nel corpus non implica che sia agrammaticale, potrebbe semplicemente riflettere casualità di campionamento o preferences stilistiche.

La soluzione ottimale richiede combinazione di evidenza corpus con giudizi sperimentali. I corpora identificano fenomeni di gradazione e generano ipotesi su fattori rilevanti; gli esperimenti testano queste ipotesi in modo controllato, manipolando sistematicamente fattori di interesse e misurando effetti sulle dipendenze variabili rilevanti.

Questioni Residue e Direzioni Future

Nonostante i progressi sostanziali nella comprensione della gradualità sintattica, questioni teoriche ed empiriche importanti rimangono aperte. Una questione fondamentale riguarda lo

status dei universali linguistici in un framework probabilistico. Se la grammatica è distribuzione di probabilità piuttosto che insieme di regole categoriche, come spiegare restrizioni universali che sembrano assolute?

Per esempio, l'Universal Grammar generativista postula che tutte le lingue hanno phrase structure configurazionale (costituenti raggruppati gerarchicamente), che operazioni sintattiche rispettano structure-dependence (non manipolano sequenze lineari arbitrarie ma solo costituenti strutturali), e che esistono solo certi tipi di regole di movimento (che rispettano località, preservano c-command, etc.).

In un modello probabilistico, tali universali potrebbero essere implementati come prior molto forti: la probabilità di grammatiche che violano questi principi è assegnata densità prossima a zero. Nessuna quantità finita di evidenza linguistica potrebbe convincere il learner ad adottare grammatica che viola structure-dependence, perché il prior bayesiano contro tale grammatica è così forte da rendere la posterior negligibile.

Questa formulazione preserva contenuto empirico degli universali (predicono che certe strutture non saranno mai osservate in nessuna lingua) pur incorporandoli in framework probabilistico. Tuttavia, solleva questione dell'origine di tali prior. Se non sono innati, come potrebbero emergere? E se sono innati, questo non reintroduce la discontinuità tra linguaggio umano e sistemi di apprendimento statistico generale?

Una possibilità è che i prior riflettano vincoli cognitivi generali non specifici al linguaggio. Per esempio, preferenza per struttura gerarchica configurazionale potrebbe derivare da architettura generale della cognizione umana, che rappresenta informazione complessa mediante strutture ricorsive gerarchiche (evidente anche in altri domini come pianificazione di azioni, categorizzazione di oggetti, ragionamento causale).

Analogamente, structure-dependence potrebbe seguire da principi di computational efficiency: regole che operano su costituenti strutturali sono più efficienti da rappresentare e applicare rispetto a regole che richiedono conteggio di posizioni lineari arbitrarie. Una grammatica che richiede "sposta il terzo elemento della frase a meno che il quinto elemento sia verbo" sarebbe computazionalmente dispendiosa e fragile (dipende da lunghezza della frase); una regola che dice "sposta il complementizzatore del clause embedded" è più semplice e robusta.

Tuttavia, queste spiegazioni funzionali rimangono speculative. Testing rigoroso richiederebbe evidenza comparativa da domini cognitivi non linguistici e da sistemi artificiali di apprendimento. Se vincoli simili emergono in reti neurali addestrate su task non linguistici ma strutturalmente analoghi, questo supporterebbe origine generale cognitiva. Se invece richiedono bias induttivi specifici al linguaggio implementati architetturalmente, questo supporterebbe qualche forma di innatismo linguistico.

Un'altra questione aperta riguarda la relazione tra gradualità sintattica e variazione sociolinguistica. Varietà diverse della stessa lingua (dialetti, socioletti) possono differire non solo in quali costruzioni sono grammaticali categorialmente, ma anche in quanto sono accettabili. Costruzioni stigmatizzate in varietà standard possono essere perfettamente naturali in varietà non-standard.

Per esempio, la doppia negazione in inglese ("I don't know nothing") è agrammaticale in standard English ma grammaticale in African American Vernacular English e molti dialetti popolari. In un framework categoriale, questo rappresenta differenza qualitativa: diverse grammatiche con regole diverse. In un framework probabilistico, potrebbe rappresentare differenza quantitativa: distribuzioni di probabilità diverse plasmate da esposizione a input diversi.

Studi sperimentali hanno mostrato che giudizi di acceptability sono modulati da esposizione recente: parlanti che ascoltano brani contenenti costruzioni particolari mostrano aumento temporaneo di acceptability di quelle costruzioni. Questo suggerisce che le rappresentazioni probabilistiche sono dinamiche, aggiornate continuamente basandosi su input. La variazione sociolinguistica potrebbe riflettere equilibri diversi determinati da pattern di esposizione diversi in comunità linguistiche diverse.

Tuttavia, alcuni fenomeni di variazione sembrano più categoriali. Per esempio, l'ordine verbo-oggetto in lingue a oggetto prima del verbo come giapponese versus verbo prima dell'oggetto in lingue come inglese non sembra questione di preferenza graduale ma di parametro strutturale. I parlanti di giapponese non giudicano ordini VO come "marginalmente accettabili"; li giudicano completamente agrammaticali. Questo suggerisce che almeno alcuni aspetti della grammatica rimangono categoriali anche in framework che ammette gradualità diffusa.

Una possibile riconciliazione è distinguere core syntax da peripheral syntax. Il core include principi strutturali astratti che sono effettivamente categoriali (almeno all'interno di una lingua/dialetto), mentre la periferia include costruzioni specifiche, collocazioni, e fenomeni di confine che mostrano gradualità. La sfida è specificare precisamente questa distinzione ed evitare che diventi puramente stipulativa.

Infine, rimane questione dell'acquisizione linguistica infantile. I bambini acquisiscono grammatica della loro lingua con velocità e successo notevoli basandosi su input relativamente limitato che include errori e enunciati incompleti. Come conciliare questo con modello probabilistico che richiede statistiche affidabili?

Una risposta è che i bambini potrebbero utilizzare prior induttivi forti che facilitano apprendimento. Anche senza regole grammaticali innate specifiche, bias generali verso certe strutture (come preferenza per gerarchie, per regolarità, per composizionalità) potrebbero accelerare convergenza su grammatica adulta. Inoltre, i bambini potrebbero essere particolarmente sensibili a pattern statistici in modi che adulti non sono, sfruttando capacità di apprendimento implicito che declina con età.

Studi di infant statistical learning hanno dimostrato che bambini di 8 mesi possono tracciare probabilità di transizione tra sillabe in stream continuo per segmentare parole, e che sensibilità a pattern distribuzionali è presente precocemente nello sviluppo. Questo supporta l'idea che apprendimento statistico è meccanismo fondamentale di acquisizione linguistica.

Tuttavia, apprendimento statistico da solo non spiega tutti i fenomeni. I bambini sembrano fare generalizzazioni che vanno oltre evidenza diretta in modi specifici. Per esempio, acquisiscono constraint su movimento wh- senza apparente evidenza negativa esplicita: nessuno dice ai bambini "non puoi estrarre da whether-island", eppure i bambini convergono

su questo constraint. Questo potrebbe riflettere prior innato contro certe strutture, oppure potrebbe emergere da pattern indiretti nell'input (per esempio, assoluta assenza di certi pattern può essere segnale informatico sufficiente se il bambino è sensibile ad esso).

La ricerca futura dovrà integrare evidenze da acquisizione linguistica, psicolinguistica, neuroscienze cognitive, linguistica computazionale, e teoria linguistica formale per sviluppare modello comprensivo della competenza linguistica che cattura sia regolarità sistematiche sia gradualità osservate empiricamente. Tale modello dovrebbe essere specifico computazionalmente, verificabile empiricamente, e cognitivamente plausibile.

L'obiettivo non è semplicemente catalogare fenomeni di gradualità, ma comprendere i principi computazionali e rappresentazionali che li generano. La dissoluzione della dicotomia grammaticale non implica abbandono di rigore teorico, ma piuttosto evoluzione verso teoria più sofisticata che integra insights da paradigmi multipli e riconosce la complessità intrinseca del sistema linguistico umano.